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

ENSAE ParisTech | École nationale de la statistique et de l'administration économique | université PARIS-SACLAY

NNT : 2017SACLG002

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École nationale de la statistique et de l'administration économique

Laboratoire d'accueil : CREST, Center for Research in Economics and Statistics, UMR 9194 CNRS

Spécialité de doctorat : Mathématiques appliquées

Vincent COTTET

Theoretical Study of some Statistical Procedures Applied to Complex Data

Date de soutenance : 17 novembre 2017

Après avis des rapporteurs :
PETER GRÜNWALD (CWI et Leiden University)
ISMAËL CASTILLO (Université Paris 6)

Jury de soutenance :

PIERRE ALQUIER	(ENSAE/CREST) Codirecteur de thèse
ISMAËL CASTILLO	(Université Paris 6) Rapporteur
OLIVIER CATONI	(CREST) Examinateur
NICOLAS CHOPIN	(ENSAE/CREST) Codirecteur de thèse
ARNAUD GUYADER	(Université Paris 6) Président
PETER GRÜNWALD	(CWI et Leiden Université) Rapporteur

Fondation mathématique FMJH Jacques Hadamard

CREST CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

PhD Thesis in Mathematics

Theoretical Study of some Statistical Procedures applied to Complex Data

Vincent COTTET

CREST, ENSAE, Université Paris Saclay

Supervisors: Pierre ALQUIER and Nicolas CHOPIN

Professors of Statistics at ENSAE

Si c'était si facile, tout le monde le ferait
Qui tu serais pour réussir où tous les autres ont échoué
Oublie tes rêves prétentieux, redescends sur terre
Ou tu n'en reviendras jamais

– Si c'était si facile, **Casseurs Flowters** –

Remerciements

Même si *le mathématicien n'est qu'une machine à transformer du café en théorème*, il n'est pas *si facile* d'en devenir une et je n'aurais pu y arriver sans l'aide et l'apport, primordial comme plus futile, de plusieurs personnes que je voudrais remercier ici.

Je remercie Nicolas, mon directeur "historique" de thèse. Dès la 2e année de l'ENSAE, il m'a encadré en stage. Le premier projet m'a permis de travailler avec Simon avec qui il a été extrêmement intéressant de collaborer. Je n'y connaissais rien, je leur suis reconnaissant de m'avoir aiguillé gentiment et patiemment dans la recherche. Ensuite, Nicolas m'a toujours fait sentir qu'il était à mes côtés tout en me laissant une entière liberté de manœuvre. Durant ces 3 ans et même si nous n'avons pas toujours travaillé directement ensemble, sa démarche scientifique, d'une rigueur et une honnêteté absolue, a été une inspiration et, disons le franchement, un modèle.

Je voudrais remercier de manière égale Pierre. Il m'a tout d'abord accueilli à Dublin puis encadré en mémoire de master. Nous avons ensuite travaillé à l'ENSAE sur deux projets, alors qu'il avait déjà une autre thèse à encadrer. J'ai sans doute beaucoup trop profité de sa porte toujours ouverte mais je ne sais pas si cette thèse aurait abouti sans cela. Il m'a montré et enseigné le travail de la recherche quasiment du début à la fin. Dans un style différent de Nicolas mais complémentaire, ses qualités intellectuelles et humaines ont été exemplaires.

Le laboratoire de statistiques du Crest m'a soutenu pendant les trois ans, me permettant de travailler dans d'excellentes conditions. Mais je retiens surtout les magnifiques personnes le composant : je me suis construit au fil des multiples discussions sur des sujets innombrables. Je remercie Guillaume, qui a accepté de collaborer avec nous, cela nous a entrainés vers des sommets insoupçonnés. Je porte également une reconnaissance éternelle à Sacha, Arnak, Anna, Olivier, Cristina et Marco. Je suis franchement fier d'avoir participé à cette équipe. Du côté de l'administration, je remercie Arnaud, Pascale et Sophie qui font le travail de l'ombre dans un environnement compliqué. En revanche, je ne suis que moyennement fier d'être un des rares diplômés de l'université Paris-Saclay ;

tout ça est assez pathétique.

Mes collègues thésards m'ont également beaucoup aidé : cela rassure de voir qu'on n'est pas tout seul dans cette galère... Et finalement, ce sont eux les plus présents au labo ! Merci à James, Pierre, Edwin, Mehdi, Alexander, Lionel, Gautier, Léna, The Tien, Mohamed et Charles.

Mathématiquement et officiellement, cette thèse n'a pas de valeur sans le travail des rapporteurs et membres du jury. Je les remercie d'avoir accepté de faire tout ce travail sans quoi la recherche ne pourrait tout simplement pas exister. Un grand merci à Ismaël, Peter, Olivier et Arnaud.

Cette thèse a bénéficié du soutien de l'ENSAE : c'est une chance énorme de pouvoir dégager autant de temps pour des travaux de recherche sur un poste Insee. Je remercie donc Julien, Romain et Lionel, qui ont préservé ces postes, ainsi que tous mes collègues de la direction des études qui ont supporté, surtout eux, ma mauvaise humeur quasi quotidienne : en premier lieu mes collègues de bureau Malika et Jérémy, mes amis économistes Vanda et Arthur et nos précieuses gestionnaires des études Mathilde et Anne ainsi qu'Emmanuelle, Stéphanie, Dorothée Romain, Claude, Patrick et le service des stages. Je remercie également tous les professeurs et chargés de TD avec qui j'ai pu travailler pendant tout ce temps. Cela a été un plaisir de travailler avec eux et m'a libéré le temps nécessaire pour ces travaux.

Si cette thèse est pour moi un accomplissement, je date le début au concours administrateur qui a été un évènement important pour moi. Je remercie toutes les personnes qui m'ont aidé à ce moment-là : Etienne, Corentin, Delphine, Xavier, Bruno.

Tous mes amis ont à un moment ou l'autre participé, peut-être sans le savoir, de près comme de loin, à ces travaux. Je souhaite remercier, entre autres : Sylvain, JF, Estelle, Brize, Claire, Isabelle, Aurélie, Marine, Gauthier, Kais et Paco.

Je veux enfin remercier ma famille pour avoir été là avant et pendant tout ce temps. Ce travail personnel n'aurait pu voir le jour sans l'environnement de stimulation intellectuelle permanent dans lequel j'ai baigné depuis tout petit grâce à mes grands-parents, mes parents, mes frères et ma sœur. Un grand merci à Julia pour ses corrections de *Ceci est MA thèse, Bordel !*. Je remercie pour finir, en quantité infinie $\times 2$, celle qui a toujours été à mes côtés dans cette aventure au long cours et sans qui tout ce travail n'aurait pas vu le jour. Bon, en fait pas mal de monde a participé à tout ça !

Enfin, je souhaite dédier cette thèse à mon grand-père Jean : même si je pense qu'il est fier de ce travail, j'espère qu'il est encore plus fier de voir que j'essaie tous les jours de suivre sa trace.

Contents

Chapitre 1

Introduction (in French)

1.1 Mise en perspective

La science des données (souvent appelée *Data Science*) a connu des grands changements récemment, que ce soit au niveau du volume des données disponibles, mais aussi du coté des capacités de traitement. Un exemple marquant est le premier séquençage du génome humain qui a été conclu en 2003. Quinze ans plus tard, le coût a été divisé par 10^5 et ne prend plus que quelques heures. En 2007, la société Netflix a organisé une compétition dont la dotation pour le vainqueur était de un million de dollars. Le but était, étant données des notes d'utilisateurs concernant des films, de prédire les notes futures. On voit ici le renversement opéré : auparavant, il était fréquent de devoir calculer un effet moyen. Les nouvelles données permettent maintenant d'individualiser les prédictions ; pour cela, il faut bien sûr développer de nouveaux outils adaptés.

Les conséquences sur les statistiques en tant que discipline scientifique ont été nombreuses. La régularisation est apparue comme une méthode intéressante pour traiter les modèles en grande dimension. Comparativement à d'autres techniques, elle a montré son intérêt tant pratique avec des algorithmes rapides que théorique où les performances sont proches de l'optimal. C'est un champ de recherche actif comme on peut le voir avec l'introduction de nouvelles régularisations telle que le SLOPE en 2015. Le cadre de la grande dimension se révèle aussi impropre aux résultats asymptotiques ; il faut alors développer de nouveaux résultats non asymptotiques qui requièrent eux-mêmes de nouvelles techniques de preuve.

En parallèle, l'augmentation de la puissance de calcul disponible a permis le développement de nouvelles méthodes d'inférence pour des modèles qu'il était impossible d'utiliser auparavant. Cela est particulièrement pour la statistique bayésienne car elle demande souvent l'usage de méthodes numériques pour calculer des estimateurs approchés. Les méthodes de MCMC (Markov Chain Monte Carlo), popularisées dans les années 1990, ont d'abord permis de traiter de un grand nombre de modèles. Récemment, les méthodes ABC (Approximate Bayesian Computation) permettent d'aborder une classe encore plus large de modèles mais cela se fait au prix d'une utilisation encore plus intense de simulations.

Cette thèse s'inscrit pleinement dans ce mouvement : les deux premiers chapitres abordent l'étude d'estimateurs pénalisés et le modèle particulier de la complétion de matrice. Le problème est apparu récemment et il s'applique particulièrement bien à des données individuelles. Nous essayons alors de faire une analyse la plus large possible : un estimateur est proposé, nous étudions théoriquement ses propriétés puis nous développons un algorithme pour enfin le tester sur des jeux de données réels. Le troisième chapitre propose une méthode approchée ABC dans le cadre bayésien qui est parallélisable. Elle est appliquée sur plusieurs modèles et montre son aptitude à être à la fois rapide et efficace.

Comme la complétion de matrice est abordée dans la majorité des chapitres, quelques rappels de l'état de l'art sont faits dans la section suivante avant de passer à la description des trois chapitres constituant le cœur de la thèse.

Notations utilisées dans ce chapitre :

- $\forall n \in \mathbb{N}^*, \quad [n] = \{1, \dots, n\}$.
- $\mathbb{R}^{m_1 \times m_2}$ sera la notation utilisée pour les matrices réelles de taille (m_1, m_2).
- Pour une matrice $M \in \mathbb{R}^{m_1 \times m_2}$, $M^\top$ est sa transposée.
- $\mathbf{e}_i^m$ représente le i-ème vecteur de la base canonique de $\mathbb{R}^m$, espace assimilé à $\mathbb{R}^{m \times 1}$. On note alors $E_{i,j} = \mathbf{e}_i^{m_1} \mathbf{e}_i^{m_2 \top}$.
- $M_{i,\cdot}$ représente la i-ème ligne de la matrice M et $M_{\cdot,j}$ représente sa j-ème colonne.
- $\sigma(M)_k$ représente la k-ème valeur singulière de M.
- Pour $p \geq 1$, on note $\|M\|_{S_p} = (\sum_{k=1}^{\min(m_1,m_2)} \sigma(M)_k^p)^{1/p}$ la norme Schatten-p de la matrice M. La norme Schatten-2 est aussi nommée *norme de Frobenius*.
- la norme d'opérateur de la matrice M est la plus grande valeur

singulière de M. Elle est notée $\|M\|_{S_\infty}$.
— pour une matrice M, on note $\|M\|_\infty$ le maximum des valeurs absolues des entrées de M.
— $\text{sign}(x) = \mathbb{1}\{x > 0\} - \mathbb{1}\{x < 0\}$.
— On pose $\mathfrak{m} = \min(m_1, m_2)$ et $\mathfrak{M} = \max(m_1, m_2)$.

1.2 Comment compléter une matrice ?

1.2.1 Cadre général et complétion exacte

La complétion de matrice est le problème consistant à deviner les entrées d'une matrice qui n'est que partiellement observée. Pour être précis, le problème n'a pas été traité dans la chronologie présentée ici, mais il semble que les travaux de Nathan Srebro avec d'autres auteurs soient pionniers (voir entre autres [Srebro et al., 2005] et [Srebro and Shraibman, 2005]). Ils utilisent la perte charnière (ou *hinge loss*) qui sera présentée plus tard.

Sans aucune hypothèse, deviner les entrées manquantes est impossible. En revanche, si la matrice a une certaine structure particulière, le problème peut devenir faisable sous certaines conditions. L'hypothèse souvent retenue et que nous utiliserons ensuite est que la matrice est de faible rang. Cette hypothèse est assez naturelle dans beaucoup d'applications : si les lignes représentent des caractéristiques d'individus empilés en colonne, le faible rang est obtenu quand les caractéristiques sont liées entre elles. Cette hypothèse est d'ailleurs à la base de l'analyse en composante principale [1].

Le faible rang : une hypothèse naturelle.

On peut ici donner un exemple qui permet de bien visualiser le problème et l'hypothèse de faible rang. Les données sont les notes données par des utilisateurs à des films qu'ils ont regardés. Il y a un très grand nombre de films et également un très grand nombre d'utilisateurs mais toutes les notes ne sont pas connues : on peut donc représenter ces données comme une matrice incomplète où les lignes sont les utilisateurs et les colonnes les films, voir le tableau 1.1.

1. Les données sont alors observées entièrement et on cherche juste une représentation de faible rang

	Le Parrain	Pulp Fiction	Rain Man	OSS117	Le Dictateur	Toy Story
Anna	5	2	.	5	.	.
Pierre	.	5	.	.	.	5
Vincent	1	.	3	5	.	.
Sophie	.	.	.	.	3	3
Jérémy	.	3	.	5	.	.
Nicolas	5	.	.	1	5	.

Tableau 1.1 – Représentation de notes de film sous forme d'une matrice incomplète

S'il était possible de prédire les notes, on pourrait alors recommander les films aux différents utilisateurs. On voit ici que les données individuelles, maintenant facilement accessibles dans plusieurs cas, permettent de faire des prévisions individualisés alors que pendant longtemps, on ne pouvait souvent que s'intéresser à la moyenne.

Le faible rang de la matrice sous-jacente s'analyse ici comme la similarité entre les colonnes : certains individus sont proches entre eux et seul un petit nombre de profils type existe. Cela peut facilement se voir en faisant une décomposition en valeur singulière (*Singular Values Decomposition*) de la matrice. Toute matrice M de taille $m_1 \times m_2$ et de rang r peut se décomposer en un produit $U\Sigma V^\top$ où :

— U est de taille $m_1 \times r$ et les colonnes sont orthonormées ;
— Σ est une matrice diagonale et les r coefficients strictement positifs $(\sigma(M)_k)_{1\leq k\leq r}$ sont les valeurs singulières ;
— V est de taille $m_2 \times r$ et les colonnes sont orthonormées.

Finalement, avec cette décomposition, on a :

$$M = \sum_{k=1}^{r} \sigma(M)_k U_{\cdot,k} (V_{\cdot,k})^\top.$$

La matrice M se décompose en une somme de r matrices de rang 1. Chaque entrée s'écrit donc :

$$M_{i,j} = U_{i,\cdot} \Sigma V_{j,\cdot}^\top = \sum_{k=1}^{r} \sigma(M)_k U_{i,k} V_{j,k}.$$

Intuitivement, la note de l'individu j pour le film i est la somme de r composantes seulement, pondérées par les valeurs singulières $\sigma(M)_k$. La

décomposition en valeurs singulières de la matrice M est lié aux diagonalisations de $MM^\top$ et de $M^\top M$: ces deux matrices partagent les mêmes valeurs propres ; les vecteurs propres de ces matrices sont respectivement les matrices U et V. Ces opérations sont à la base de l'analyse en composantes principales (ACP), qui peut donc être vue comme une SVD de la matrice des observations. Enfin, la SVD présentée ici est réduite et peut être augmentée avec des valeurs singulières nulles et des vecteurs singuliers associés.

Conditions pour une complétion exacte.

Une question qui a intéressé la communauté consiste en l'étude des conditions permettant la complétion exacte. Ici, M sera une matrice carrée de taille m (le cas le plus difficile) de rang r ; on note $U\Sigma V^\top$ sa décomposition en valeurs singulières. Les entrées observées sont représentées par n couples à valeur dans $[m] \times [m]$; ils constituent l'ensemble Ω. À partir de la décomposition en valeurs singulières, on voit que M a $2mr - r^2$ degrés de liberté : la matrice Σ a r entrées non nulles, les vecteurs singuliers à gauche ont $(m-1) + \cdots + (m-r) = mr - r(r+1)/2$ degrés de liberté et autant pour les vecteurs à droite. Si $n < 2mr - r^2$, plusieurs matrices peuvent correspondre et il n'y a pas unicité de la solution.

On voit alors que $2mr - r^2$ est la borne minimale. La question naturelle qui apparaît est alors de savoir quel est la taille de l'échantillon qu'il faut quand les entrées sont tirées au hasard car cela doit être supérieur à la borne inférieure. Il est naturel de considérer en première approche le tirage uniforme : chaque échantillon de taille n parmi les m^2 entrées est équiprobable.

Par ailleurs, il est facile de voir que certaines matrices sont plus difficiles à reconstruire que d'autres. Par exemple, une matrice ayant des 0 partout sauf à quelques endroits est très difficile à reconstruire car on ne sait pas où les entrées non nulles se trouvent. Pour la matrice

$$M = \mathbf{e_1^m} \mathbf{e_m^{m\top}} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{1.1}$$

l'entrée en haut à droite doit se trouver dans l'échantillon observé sinon elle est impossible à deviner. Si les observations proviennent d'un tirage uniforme, la probabilité est de n/m^2, ce qui est usuellement faible car n

est petit devant m^2. Les conditions portent alors sur les vecteurs singuliers de M, qui ne doivent pas être trop piquées et dont les entrées doivent être bien réparties. La matrice de l'exemple (1.1) a des vecteurs singuliers avec un seul 1 et des 0 sinon, ce qui correspond au cas compliqué.

La relaxation du rang.

Ceci n'est pas du jargon militaire, mais a trait au programme de reconstruction. Intuitivement, il y a un très grand nombre de matrices correspondant aux entrées observées. Pour en trouver une de faible rang, le programme naturel est alors :

$$\begin{array}{rl} \text{minimiser} & \operatorname{rang}(X) \\ \text{sous la contrainte} & \forall (i,j) \in \Omega,\ X_{i,j} = M_{i,j} \end{array} \tag{1.2}$$

La difficulté principale de ce programme est que la fonction à minimiser dans (1.2) n'est pas convexe. Une idée fondamentale est alors de remplacer le rang par la norme nucléaire. La norme nucléaire, aussi appelée norme Schatten-1 est la somme des valeurs singulières de la matrice M :

$$\|M\|_{S_1} = \sum_{k=1}^{r} \sigma(M)_k.$$

Comme une norme est convexe, le programme suivant :

$$\begin{array}{rl} \text{minimiser} & \|X\|_{S_1} \\ \text{sous la contrainte} & \forall (i,j) \in \Omega,\ X_{i,j} = M_{i,j} \end{array} \tag{1.3}$$

est alors beaucoup plus simple à traiter. C'est d'ailleurs un programme d'optimisation SDP (*Semidefinite Programing*). À notre connaissance, ce programme est proposé pour la première fois par Fazel et al. [2001]. La norme nucléaire a été ensuite largement adoptée pour plusieurs problèmes similaires.

Nous avons maintenant tous les ingrédients pour énoncer les théorèmes principaux de la complétion exacte.

Théorème 1.1 (Théorème 1.2 de [Candès and Tao, 2010])**.** *Soit M une matrice de taille $m_1 \times m_2$, de rang r observant la* strong incoherence property [2] *de paramètre μ et notons $\mathfrak{M} = \max(m_1, m_2)$. Supposons qu'on*

2. la définition précise de cette condition est dans l'article. Le but est d'éliminer les cas pathologiques comme celui présenté en (1.1).

observe n entrées tirées au hasard selon une loi uniforme. Il existe alors une constante absolue C telle que si :

$$n \geq C\mu^2\, \mathfrak{M}\, r(\log \mathfrak{M})^6,$$

alors M est l'unique solution du programme (1.3) *avec une probabilité plus grande que $1-n^{-3}$. Autrement dit : avec grande probabilité, la minimisation de la norme nucléaire permet de retrouver la matrice d'origine.*

Ce résultat est une amélioration d'un article précédent [Candès and Recht, 2012] qui a un résultat où le nombre minimal d'observations est de l'ordre de $\mathfrak{M}^{1.2}\, r \log \mathfrak{M}$ pour permettre de reconstruire avec grande probabilité la matrice. C'est un résultat important et surprenant : à un terme logarithmique, il suffit ainsi de l'ordre de $\mathfrak{M}\, r$ entrées connues pour reconstruire toute la matrice, ce qui est bien plus que faible que $\mathfrak{M}^2$ si le rang est petit. On est ainsi proche du nombre minimal d'entrées à connaitre dans le meilleur des cas pour un tirage aléatoire uniforme des entrées.

1.2.2 Complétion en présence de bruit par moindres carrés

Le problème posé précédemment est assez particulier car l'aléatoire ne se trouvait que dans le tirage des entrées observées. La question de la reconstruction exacte peut paraître excessif dans le sens où, pour de nombreuses applications, il est suffisant d'avoir une prévision ayant une marge d'erreur faible. Deuxièmement, l'observation sans bruit est peu réaliste. Dans le cas bruité, on se trouve alors dans un cadre plus classique pour le statisticien où il y a un paramètre à estimer. Dans la suite, la matrice à reconstruire sera de taille $m_1 \times m_2$.

Modèle de régression *Trace*.

Nous introduisons ici le modèle de régression *trace*, qui a un cadre plus général que la complétion de matrice. On munit l'ensemble des matrices réelles de taille $m_1 \times m_2$ du produit scalaire canonique :

$$\langle A, B\rangle = \mathrm{Tr}(A^\top B) = \mathrm{Tr}(B^\top A).$$

La norme associée est la norme de Frobenius, qui est aussi la norme Schatten-2 :

$$\|A\|_{S_2} = \sqrt{\mathrm{Tr}(A^\top A)} = \sqrt{\sum_{i,j} A_{i,j}^2}.$$

Le cadre statistique est alors le suivant : pour $i \in \{1, \dots, n\}$, on observe $Y_i \in \mathbb{R}$ et $X_i \in \mathbb{R}^{m_1 \times m_2}$ suivant le modèle :

$$Y_i = \langle M^\star, X_i \rangle + \sigma \varepsilon_i,$$

où $M^\star$ est le paramètre à estimer et (ε_i) une suite indépendante et identiquement distribuée de bruit centré et réduit.

La complétion de matrice est alors un cas particulier où les matrices de design (X_i) sont des matrices *masque* : elles sont nulles partout sauf en une entrée qui vaut 1 ; elles sont donc à valeur dans :

$$\mathcal{X} = \left\{ \mathbf{e}_j^{m_1 \top} \mathbf{e}_k^{m_2},\ (j,k) \in [m_1] \times [m_2] \right\}.$$

Dans ce cas, si l'entrée (k,l) est non nulle donc si $X_i = E_{k,l}$, on a pour toute matrice M : $\langle M, X_i \rangle = M_{k,l}$.

L'estimateur couramment proposé est, pour $\lambda \in \mathbb{R}$, l'estimateur des moindres carrés pénalisés :

$$\widehat{M} = \operatorname*{arg\,min}_{M \in \mathcal{C}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle M, X_i \rangle)^2 + \lambda \left\| M \right\|_{S_1} \right\}, \tag{1.4}$$

où $\mathcal{C}$ est un sous-ensemble, habituellement convexe, de $\mathbb{R}^{m_1 \times m_2}$.

Plusieurs travaux se sont intéressés aux propriétés théoriques de cette estimateur en fonction de différentes conditions sur le bruit, le design et la matrice $M^\star$. Après les travaux [Bach, 2008], [Candès and Plan, 2010], [Keshavan et al., 2010], [Rohde and Tsybakov, 2011] et [Negahban and Wainwright, 2012] (entre autres), nous citerons ici un résultat de [Klopp, 2014] qui traite un cadre assez général.

Résultat principal de [Klopp, 2014].

Deux conditions portent sur le tirage des entrées. Il faut que les entrées ainsi que les lignes et les colonnes aient toutes une probabilité pas trop petite d'être tirées : on peut s'écarter du tirage uniforme mais sans aller vers des cas trop extrêmes. Les matrices X_i sont tirées indépendamment et selon la même loi Π dans $\mathcal{X}$. Si on note $\pi_{j,k}$ la probabilité que X_i soit égale à $E_{j,k}$, le théorème principal a besoin de deux conditions :

Hypothèse 1.1. *Il existe une constante $L > 1$ telle que :*

$$\left. \begin{matrix} \max_j \left(\sum_{k=1}^{m_2} \pi_{j,k} \right) \\ \max_k \left(\sum_{j=1}^{m_1} \pi_{j,k} \right) \end{matrix} \right\} \leq L / \mathfrak{m}\,.$$

Hypothèse 1.2. *Il existe une constante $\mu \geq 1$ telle que, pour toute localisation (j,k) :*

$$\pi_{j,k} \geq (\mu m_1 m_2)^{-1}.$$

Dans le cas d'un tirage uniforme, on a $L = \mu = 1$. La dernière hypothèse se rapporte au bruit.

Hypothèse 1.3. *Il existe $K > 0$ tel que :*

$$\max_{i \in \{1,\dots,n\}} \mathbb{E} \exp\left(|\varepsilon_i| / K\right) < \infty.$$

Il est alors possible d'énoncer le théorème principal.

Théorème 1.2 (théorème 7 de [Klopp, 2014])**.** *Soit (X_i) un tirage i.i.d. de loi Π sur $\mathcal{X}$ satisfaisant les hypothèses 1.1 et 1.2. Soit $M^\star$ une matrice réelle telle que $\|M\|_\infty \leq \boldsymbol{a}$ et que le bruit satisfasse l'hypothèse 1.3. On prend :*

$$\lambda = C\sigma\sqrt{\frac{L \log(m_1 + m_2)}{n\, \mathfrak{m}}},$$

où C est une constante connue. On considère l'estimateur $\widehat{M}$ satisfaisant (1.4) *où $\mathcal{C} = \{M : \|M\|_\infty \leq \boldsymbol{a}\}$. Il existe alors une constante c' dépendant uniquement de $(K, \sigma, \boldsymbol{a}, \mu, L)$ telle que :*

$$\frac{\left\|\widehat{M} - M^\star\right\|_{S_2}^2}{m_1 m_2} \leq c' \max\left\{\frac{\operatorname{rang}(M^\star)\, \mathfrak{M} \log(m_1 + m_2)}{n}, \sqrt{\frac{\log(m_1 + m_2)}{n}}\right\}.$$

Le parallèle avec le Lasso est clair. La norme nucléaire remplace la norme ℓ_1, et est d'ailleurs la norme ℓ_1 appliquée aux valeurs singulières. La taille du paramètre est de l'ordre de $r\,\mathfrak{M}$ et la borne supérieure est de l'ordre de $r\,\mathfrak{M}/n$ à des termes logarithmiques près. Nous verrons dans le chapitre 4 qu'un cadre unifié existe, en suivant des idées de [Lecué and Mendelson, 2015a,b].

Différents résultats de bornes inférieures établissent une vitesse minimax de l'ordre de $r\,\mathfrak{M}/n$. La borne supérieure donnée précédemment est donc optimale à un facteur logarithmique près. La dépendance du

résultat au terme de bruit peut-être éliminée en considérant un autre estimateur similaire au *square-root Lasso*, en prenant la racine carrée du terme d'ajustement. Cet estimateur est étudié dans le même article. En pratique, on utilise la validation croisée pour ajuster λ. Dans un modèle légèrement différent, la même auteure arrive à enlever le log et obtient la vitesse minimax, voir [Klopp, 2015].

L'article [Koltchinskii et al., 2011] traite un cadre plus général de régression trace et pas seulement avec des matrices masque. L'estimateur proposé est alors un peu différent et si le tirage est non-uniforme, il doit être connu.

1.2.3 Quelques algorithmes

Il nous semble important d'aborder ici le problème du calcul des estimateurs même si ce n'est pas au cœur des travaux qui suivent. La fonction à minimiser en (1.4) est convexe, ce qui est intéressant, mais il reste des questions importantes. Nous allons commencer par rappeler deux propriétés importantes de la norme nucléaire. Le sous-différentiel de la norme nucléaire (voir [Watson, 1992]) en M dont la SVD est $U\Sigma V^\top$ est :

$$\partial \|\cdot\|_{S_1}(M) = \left\{UV^\top + W, \quad \|W\|_{S_\infty} \leq 1,\, U^\top W = 0, WV = 0\right\}.$$

Quand la matrice M est de faible rang, le sous-différentiel de la norme nucléaire est grand car il y a beaucoup de possibilités pour W. Une fonction qui revient souvent est la fonction de seuillage doux des valeurs singulières. On la définit par :

$$S_\lambda(M) = UDV^\top \quad \text{où} \quad D = \operatorname{diag}\left(\{(\max(0, \sigma_k(M) - \lambda)\}_{1\leq k\leq r}\right).$$

Comme pour les vecteurs, on a alors le résultat entre l'opérateur proximal de la norme nucléaire et le seuillage doux des valeurs singulières S_λ.

Proposition 1.1. *Soit M une matrice de $\mathbb{R}^{m_1\times m_2}$. Alors :*

$$\arg\min_X \left\{\frac{1}{2}\|X - M\|_{S_2}^2 + \lambda\|X\|_{S_1}\right\} \ni S_\lambda(M).$$

Démonstration, en suivant [Recht et al., 2010]. On pose la fonction h, définie pour toute matrice X par $h(X) = \frac{1}{2}\|X - M\|_{S_2}^2 + \lambda\|X\|_{S_1}$. La

fonction h est strictement convexe car somme de deux fonctions strictement convexes, donc le minimum est unique. X_0 est le minimiseur de h si :

$$0 \in \partial h(X_0) = \left\{ X_0 - M + \lambda D \, : D \in \partial \left\| \cdot \right\|_{S_1} (X_0) \right\}.$$

On va montrer que la relation précédente est vraie pour $X_0 = S_\lambda(M)$. On note U_0, V_0 les vecteurs singuliers associés aux valeurs singulières de M supérieures à λ et U_1, V_1 pour les valeurs inférieures, Σ_0 et Σ_1 désignant les deux matrices diagonales correspondantes. On a alors :

$$X_0 = S_\lambda(M) = U_0(\Sigma_0 - \lambda I)V_0^\top.$$

On pose alors, pour que tout marche bien :

$$W_0 = \lambda^{-1} U_1 \Sigma_1 V_1^\top.$$

On voit que $\left\| \lambda^{-1} U_1 \Sigma_1 V_1^\top \right\|_{S_\infty} \leq 1$ car les entrées de Σ_1 sont bornées par λ. En outre, $U_0^\top W_0 = W_0 V_0 = 0$ car les vecteurs de U_1 sont orthogonaux à ceux de U_0 et pareillement pour V_1 et V_0. Finalement, $U_0 V_0^\top + W_0$ est dans $\partial \left\| \cdot \right\|_{S_1} (M_0)$.

Après avoir exhibé un astucieux élément du sous-différentiel de la norme nucléaire en X_0, on peut maintenant conclure car :

$$X_0 - M + \lambda \left(U_0 V_0^\top + W_0 \right) = 0.$$

■

La proposition 1.1 traite finalement un cas proche de la complétion de matrice, qui peut plutôt se voir comme un problème de débruitage : toutes les entrées sont observées et on cherche une représentation de faible rang qui colle aux données [3]. Cet opérateur proximal est à la base de beaucoup d'algorithmes traitant la complétion de matrice, le premier d'une longue liste étant [Cai et al., 2010]. L'opérateur proximal S_λ est aussi utilisé dans les algorithmes de type *ADMM* (Alternating Direction Method of Multipliers), qui sont aussi populaires en complétion de matrice.

Nous en proposons un (algorithme 1.1) qui suit [Boyd et al., 2011] dans la forme réduite. Le but est de calculer le minimiseur sans la contrainte :

$$\widehat{M} = \underset{M \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle M, X_i \rangle)^2 + \lambda \left\| M \right\|_{S_1} \right\},$$

Algorithme 1.1 Algorithme ADMM pour la complétion de matrice

Initialisation $\varepsilon, M^0, N^0, U^0, t = 0$.
Tant que $\left\|U^t - U^{t-1}\right\|_{S_2} > \varepsilon$, faire :

1. $t \leftarrow t + 1$
2. $M^t \leftarrow \arg\min_M \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \langle M, X_i \rangle)^2 + \frac{\rho}{2} \left\| M - N^{t-1} + U^{t-1} \right\|_{S_2} \right\}$
3. $N^t \leftarrow \arg\min_N \left\{ \lambda \|N\|_{S_1} + \frac{\rho}{2} \left\| M^t - N + U^{t-1} \right\|_{S_2} \right\}$
4. $U^t \leftarrow U^{t-1} + M^t - N^t$

Retourner M^t.

en séparant le problème en deux parties identifiées par deux matrices M et N dans l'algorithme.

Dans cet algorithm, la mise à jour de M se fait entrée par entrée et est peut coûteuse. En revanche, la mise à jour de N implique une SVD en utilisant l'opérateur proximal. Le problème de ces algorithmes est que le calcul de la SVD est très coûteux, de l'ordre de $\mathfrak{M}\mathfrak{m}^2$. Différentes méthodes sont alors apparues pour contourner le problème.

La première, et la plus simple, est de calculer une SVD approchée : on ne calcule par exactement les valeurs et les vecteurs singuliers, en s'autorisant une certaine marge ; il existe des méthodes stochastiques rapides. Dans le même genre d'idées, certains algorithmes autorisent le calcul des N plus grandes valeurs propres, N étant fixé avant l'exécution de l'algorithme. Comme S_λ ne requiert que le calcul des valeurs singulières supérieures à λ, on peut fixer N petit et l'augmenter progressivement si la plus petite valeur singulière est plus grande que λ. Cette méthode est assez difficile à calibrer mais les routines calculant les N plus grandes valeurs singulières sont implémentées en FORTRAN.

Un problème bi-convexe associé.

Les méthodes de SVD approchées ne sont pas très satisfaisantes car elles sont très difficiles à calibrer. De plus, le besoin de mémoire est très important car des matrices de tailles $m_1 \times m_2$ sont stockées : on ne profite pas de l'approximation de faible rang. On peut alors utiliser une

3. L'ACP correspondrait alors à l'opérateur de seuillage dur pour le même problème.

caractérisation de la norme nucléaire utilisant une factorisation :

$$\forall M \in \mathbb{R}^{m_1 \times m_2}, \quad \|M\|_{S_1} = \frac{1}{2} \min_{\substack{LR^\top = M \\ L \in \mathbb{R}^{m_1 \times \mathfrak{m}} \\ R \in \mathbb{R}^{m_1 \times \mathfrak{m}}}} \left\{ \|L\|_{S_2}^2 + \|R\|_{S_2}^2 \right\}.$$

En utilisant la SVD de $M = U\Sigma V^\top$ et en posant $L_0 = U\Sigma^{1/2}, R_0 = V\Sigma^{1/2}$, on a immédiatement $\|X\|_{S_1} = \|L_0\|_{S_2}^2 + \|R_0\|_{S_2}^2$. Pour montrer l'autre sens de l'inégalité, on peut utiliser la caractérisation *Semidefinite Programming* de la norme nucléaire. Par une double minimisation, on a alors que l'estimateur (1.4) est alors identique à $\widehat{L}\widehat{R^\top}$ où :

$$(\widehat{L}, \widehat{R}) = \operatorname*{arg\,min}_{\substack{L \in \mathbb{R}^{m_1 \times \mathfrak{m}} \\ R \in \mathbb{R}^{m_1 \times \mathfrak{m}}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \left\langle LR^\top, X_i \right\rangle)^2 + \frac{\lambda}{2} \left[\|L\|_{S_2}^2 + \|R\|_{S_2}^2 \right] \right\}.$$

Ce programme est beaucoup plus sympathique car il n'implique que des termes quadratiques, facilement différentiables. De plus, le problème est convexe en L et convexe en R (d'où le terme de biconvexe). Malheureusement, il n'est pas convexe en le couple. Cela veut dire qu'il y a potentiellement des minimums locaux. Pourtant, en pratique, il a été popularisé par [Mazumder et al., 2010] et une version parallélisable se trouve dans [Recht and Ré, 2013]. L'article théorique [Burer and Monteiro, 2003] montre que le problème n'est pas si mal posé et qu'en fait, le problème non convexe n'aurait qu'une solution unique. C'est donc une piste prometteuse pour développer des algorithmes rapides.

1.2.4 La complétion de matrice binaire

Après le problème de complétion de matrice, où était utilisée majoritairement la perte quadratique, les statisticiens se sont intéressés à d'autres problèmes proches. Par exemple, on peut se poser la question de la prédiction quand les entrées sont binaires. Cela arrive fréquemment sur diverses plateformes internet où l'utilisateur a le choix entre deux choix comme *j'aime* et *je n'aime pas*. Ce cas, qu'on peut appeler la *classification* par rapport au problème de *régression* étudié précedemment appelle de nouveaux outils. On labellisera les entrées observées par $\{-1, +1\}$.

Une modélisation courante pour traiter le cas de classification est de postuler un modèle de régression généralisée. L'interprétation de ce modèle est facilitée avec l'incorporation d'une variable latente $Y_i^\star$. Celle-ci peut s'interpréter, dans le cas des films, comme une mesure continue du

goût d'un utilisateur pour un film. L'observation est alors $+1$ si ce goût est supérieur à un seuil, normalisé à 0, et -1 sinon. Mathématiquement cela s'exprime comme :

$$Y_i^\star = \langle M^\star, X_i \rangle + Z_i,$$

où Z_i a pour fonction de répartition F symétrique. On a alors :

$$Y_i = \mathbb{1}\{Y_i^\star \geq 0\} - \mathbb{1}\{Y_i^\star < 0\}.$$

On peut alors directement calculer la loi de Y_i :

$$Y_i = \begin{cases} +1 & \text{avec probabilité } \mathbb{P}(Z_i \geq -\langle M^\star, X_i \rangle) = F(\langle M^\star, X_i \rangle) \\ -1 & \text{avec probabilité } \mathbb{P}(Z_i < -\langle M^\star, X_i \rangle) = F(-\langle M^\star, X_i \rangle) \end{cases} \tag{1.5}$$

La log vraisemblance normalisée, qui sera utilisée comme terme d'ajustement, est :

$$L(M) = \frac{1}{n} \sum_{i=1}^{n} \log F(Y_i \langle M, X_i \rangle).$$

Les choix usuels pour les fonctions F sont :
— la fonction logistique, $F(x) = (1 + \exp(-x))^{-1}$
— la fonction de répartition d'une loi normale centrée réduite, habituellement notée Φ.

Ce modèle a été étudié par plusieurs auteurs. Le premier travail [Davenport et al., 2014] pose le modèle de régression binaire. Le tirage des entrées est uniforme [4] et l'estimateur proposé est solution du programme suivant :

$$\begin{aligned} \text{maximiser} \quad & L(M) \\ \text{sous la constrainte} \quad & \|M\|_{S_1} \leq \alpha\sqrt{sm_1m_2} \text{ et } \|M\|_\infty \leq \alpha. \end{aligned} \tag{1.6}$$

Ici, s n'est pas le rang mais un paramètre très fortement relié à la norme nucléaire. On voit d'ailleurs qu'il est nécessaire de connaître s pour calculer l'estimateur. Les résultats obtenus dans le papier ne concernent alors pas la reconstruction de matrices de faible rang mais des matrices ayant une norme nucléaire bornée. Les conditions portant sur F sont techniques et portent sur la régularité de la fonction. Le résultat portant sur la reconstruction de $M^\star$ est alors le suivant.

4. Formellement, c'est un modèle différent, similaire dans l'esprit de celui dans [Klopp, 2015].

Théorème 1.3 (théorème 1 de [Davenport et al., 2014]). *Supposons que* $\|M^\star\|_{S_1} \leq \alpha\sqrt{sm_1m_2}$ *et* $\|M^\star\|_\infty \leq \alpha$. *Considérons les observations générées suivant le modèle* (1.5) *avec un échantillonnage uniforme. Soit l'estimateur* $\widehat{M}$ *calculé d'après* (1.6). *Alors, si* $n \geq (m_1+m_2)\log(m_1m_2)$, *avec grande probabilité, on a :*

$$\frac{1}{m_1m_2}\left\|\widehat{M} - M^\star\right\|_{S_2}^2 \leq C\sqrt{\frac{s(m_1+m_2)}{n}},$$

où C *est une constante dépendant de* F *et* α.

De prime abord, la vitesses est en racine carrée, ce qui semble plutôt mauvais. Cela vient du fait que la classe de matrices considérée (les matrices dont la norme nucléaire est bornée) est plus large que la classe des matrices de faible rang. Dans cette classe, la vitesse est optimale (nous la retrouverons d'ailleurs dans le chapitre 4). Le problème ici est qu'on ne retrouve pas la sparsité induite par la régularisation de la norme nucléaire.

Ce travail est fait dans une série d'articles [Lafond et al., 2014] [Klopp et al., 2015]. Le lien entre la norme nucléaire et le rang est alors fait et la vitesse est la vitesse obtenue pour la reconstruction avec la perte quadratique.

On rappelle ici le théorème principal, en omettant les conditions techniques sur la régularité de la fonction F. L'estimateur est alors l'estimateur où la log-vraisemblance remplace le terme d'ajustement quadratique :

$$\widehat{M} = \underset{\|M\|_\infty < \mathbf{a}}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\log F(Y_i\langle M, X_i\rangle) + \lambda\|M\|_{S_1}\right\} \tag{1.7}$$

Théorème 1.4 (Corollaire 2 de Klopp et al. [2015]). *Prenons des observations* $(X_i, Y_i)_{1\leq i\leq n}$ *i.i.d issues du modèle* (1.5) *et où* (X_i) *est un tirage de loi* Π *sur* $\mathcal{X}$ *satisfaisant les hypothèses 1.1 et 1.2. Soit* $M^\star$ *une matrice réelle telle que* $\|M^\star\|_\infty \leq \boldsymbol{a}$. *On prend :*

$$\lambda = C\sigma\sqrt{\frac{\log(m_1+m_2)}{\mathfrak{m}\,n}},$$

où C *est une constante connue dépendant de* L *et* F. *On considère l'estimateur* $\widehat{M}$ *satisfaisant* (1.7). *Il existe alors une constante* c' *dépendant*

uniquement de $(F, \boldsymbol{a}, \mu, L)$ *telle que :*

$$\frac{\left\|\widehat{M} - M^\star\right\|_{S_2}^2}{m_1 m_2} \leq c' \max\left\{\frac{\mathrm{rang}(M^\star)\,\mathfrak{M}\log(m_1+m_2)}{n}, \sqrt{\frac{\log(m_1+m_2)}{n}}\right\},$$

avec probabilité plus grande que $1 - 3/(m_1 + m_2)$.

On retrouve alors la vitesse usuelle de l'ordre $\mathrm{rang}(M^\star)\,\mathfrak{M}\log(m_1 + m_2)/n$. Les auteurs exhibent une borne inférieure de l'ordre $\mathrm{rang}(M^\star)\,\mathfrak{M}\,/n$, qui rencontre donc la borne supérieure du risque de l'estimateur à un facteur logarithmique près.

Un autre point de vue sur la complétion de matrice binaire.

Si ce résultat sur la reconstruction de $M^\star$ est optimal, il postule le modèle de régression binaire. Plus important, la reconstruction de $M^\star$ est-elle le critère le plus intéressant ? Dans le cadre de la classification, on peut supposer que le plus important est de contrôler la proportion de mauvaise prédiction. Formellement, pour un tirage X_i, le risque en prévision 0/1 pour une matrice M est alors :

$$R_{0/1}(M) = \mathbb{P}[\mathrm{sign}(\langle M, X_i\rangle) \neq Y_i] = \mathbb{E}[\mathbb{1}\{\mathrm{sign}(\langle M, X_i\rangle) \neq Y_i\}]. \quad (1.8)$$

Une étude intéressante et complète pour différentes fonctions de pertes se trouve dans [Zhang, 2004] : cet article fait le lien entre le risque des fonctions utilisées en pratique, qui sont habituellement des substituts convexes de la perte 0/1 et le risque 0/1. Cet article montre que, si la classification est presque optimale avec la perte logistique, cela ne se transfère à la perte 0/1 qu' à la puissance 1/2. Il est donc naturel d'étudier d'autres fonctions de perte qui permettraient alors d'atteindre des performances optimales pour le risque 0/1. C'est l'objet des chapitres 3 et 4.

1.3 Le cadre Bayésien - Complétion par perte quadratique

On peut légitimement appeler le cadre de travail précédent *fréquentiste* d'un point de vue statistique. Le cadre bayésien s'applique aussi à ce problème mais on doit s'y attaquer légèrement différemment. Nous

présentons ici le cadre appelé PAC-Bayésien plutôt que le cadre bayésien habituel car c'est celui qui sera repris dans le chapitre 3. Les différences sont minimes et concernent plutôt la présentation du modèle. Grossièrement, en notant θ le paramètre, l'estimation PAC-Bayésienne se construit à partir d'une fonction de perte empirique (qui remplace la log-vraisemblance), notée [5] $r(\theta)$, et une loi a priori sur θ, notée $\pi(d\theta)$. On définit alors la loi pseudo a posteriori par :

$$\widehat{\rho}_\tau(d\theta) = \frac{\exp[-\tau r(\theta)]}{\int \exp[-\tau r] d\pi} \pi(d\theta). \tag{1.9}$$

Le nombre positif τ est appelé la température inverse et joue le rôle opposé à λ dans l'estimateur pénalisé (1.4). L'estimateur PAC-Bayésien est alors calculé comme l'espérance par rapport à cette loi pseudo posterior :

$$\widehat{\theta}_\tau = \int \theta \widehat{\rho}_\tau(d\theta).$$

Pour faire le parallèle avec les estimateurs proposés précédemment, la loi a priori joue le rôle de régularisation en favorisant les paramètres là où la loi est chargée. L'estimateur, plutôt qu'être ponctuel, est une moyenne suivant la pseudo-loi a posteriori. La fonction de perte utilisée est, pour la complétion de matrice usuelle, la fonction de perte quadratique : $r(M) = 1/n \sum_{i=1}^{n} (Y_i - \langle X_i, M \rangle)^2$. Cela correspond à un bruit gaussien.

Une difficulté ici est de construire une loi a priori sur l'espace des matrices qui favorise les matrices de faible rang. Pour cela et en suivant les premiers articles bayésiens [Lim and Teh, 2007, Salakhutdinov and Mnih, 2008], on utilise une factorisation en deux matrices. En effet, si M est une matrice de rang inférieur à $K \leq \mathfrak{m}$, il existe deux matrices $L \in \mathbb{R}^{m_1 \times K}$ et $R \in \mathbb{R}^{m_2 \times K}$ telles que :

$$M = LR^\top.$$

Cette décomposition n'est pas unique. Si le rang r de M est strictement inférieur à K, les facteurs pourront avoir exactement r colonnes non nulles. Inversement, les couples (L, R) permettent de générer n'importe quelle matrice de rang inférieur à K. Le choix de K est à faire par l'utilisateur en pratique, mais, pour la théorie, on peut prendre n'importe quelle valeur et en particulier la plus générale qui est $\mathfrak{m}$. L'important est que les résultats soient adaptatifs et qu'ils ne dépendent pas de K : c'est

5. la dépendance de $r(\theta)$ aux données est implicite ici.

le cas des résultats suivants. Une grande valeur implique de stocker un paramètre plus grand mais permet de reconstruire des matrices de rang plus important. Dans la suite, on prend $K = \mathfrak{m}$ même si en pratique, il est usuel de prendre une valeur beaucoup plus petite.

Le faible rang sera, dans ce cadre, induit par la loi a priori qui doit favoriser les colonnes entièrement nulles. Ces idées sont assez similaires au *Group-LASSO* introduit par Yuan and Lin [2006] ; un modèle bayésien est étudié dans [Kyung et al., 2010]. L'idée est alors de prendre une loi a priori hiérarchique avec un paramètre additionnel $\gamma = (\gamma_k)_{1 \leq k \leq \mathfrak{m}}$. La loi a priori sur les entrées de L et de R est une loi centrée en 0. La dispersion, indexée par γ_k, va dépendre de la colonne : en effet, le but est que la colonne entière soit nulle. Le modèle hiérarchique permet cette souplesse en autorisant certaines colonnes à avoir une grande dispersion et à d'autres d'avoir une très faible dispersion et donc en annulant presque complètement la colonne en question. Formellement, voici la loi a priori usuellement utilisée :

$$\begin{aligned} \forall (i,k) \in [m_1] \times [\mathfrak{m}], \quad & L_{i,k} | \gamma_k \underset{i.i.d}{\sim} \mathcal{N}(0, \gamma_k) \\ \forall (j,k) \in [m_2] \times [\mathfrak{m}], \quad & R_{j,k} | \gamma_k \underset{i.i.d}{\sim} \mathcal{N}(0, \gamma_k) \\ \forall k \in [\mathfrak{m}], \quad & \gamma_k \underset{i.i.d}{\sim} \pi^\gamma \end{aligned} \tag{1.10}$$

où π^γ est une loi à valeur dans $\mathbb{R}^+$ qui sera précisée en temps utile. Le modèle conjugué est obtenu en utilisant la loi inverse-gamma comme habituellement pour la variance d'une loi normale.

Le paramètre finalement utilisé sera : $\theta = (L, R, \gamma) \in (\mathbb{R}^{m_1 \times \mathfrak{m}}) \times (\mathbb{R}^{m_2 \times \mathfrak{m}}) \times \mathbb{R}^{\mathfrak{m}}$. La matrice estimée sera alors :

$$\widehat{M}_\tau = \int L R^\top \widehat{\rho}_\tau(d\theta).$$

On peut alors s'intéresser aux propriétés de cette matrice en la comparant à l'oracle $M^\star$. Ce sera une approche *machine learning* dans le sens où on ne suppose pas un modèle statistique. Ensuite, nous étudierons le problème du calcul de $\widehat{\theta}_\tau$: en effet, la loi a posteriori n'a pas de formule explicite, comme fréquemment en bayésien quand le modèle est un peu complexe. On verra alors deux méthodes pour approcher l'estimateur bayésien.

1.3.1 Garanties théoriques d'un estimateur bayésien

The Tien Mai et Pierre Alquier ont étudié les propriétés de l'estimateur PAC-bayésien dans [Mai and Alquier, 2015]. Nous présentons ici leur principal résultat. On fait l'hypothèse que les données ont été générées suivant le modèle suivant :

$$Y_i = \langle M^\star, X_i \rangle + \varepsilon_i, \tag{1.11}$$

où $(\varepsilon_i)_{1\leq i\leq n}$ est séquence i.i.d. de bruit centré et les $(X_i)_{1\leq i\leq n}$ sont i.i.d. à valeur dans $\mathcal{X}$ suivant la loi Π. La connaissance de la distribution du bruit n'est pas nécessaire ici. Le but est alors de contrôler l'écart entre $M^\star$ et $\widehat{M}_\tau$. Les auteurs pour cela regardent la norme de Frobenius pondérée :

$$R(M) = \|M - M^\star\|^2_{S_2,\Pi} = \sum_{j,k} \pi_{j,k}(M - M^\star)^2_{j,k} = \mathbb{E}[\langle X_i, M - M^\star\rangle^2].$$

On ne fait pas ici d'hypothèse sur le tirage des entrées observées c'est à dire sur la loi Π. Si l'hypothèse 1.2 est vérifiée pour un certain $\mu \geq 1$, donc si chaque entrée a une probabilité minorée d'être choisie, on a alors l'inégalité reliant les deux normes :

$$\|M - M^\star\|^2_{S_2,\Pi} \geq \frac{1}{\mu}\frac{1}{m_1 m_2}\|M - M^\star\|^2_{S_2}.$$

Pour des raisons techniques, la loi a priori sur les entrées de L et de R doit être bornée : les auteurs utilisent alors une loi très simple, uniformes en fonction de *l'activation* ou non de la colonne (on fixe $\delta \gg \kappa$) :

$$L_{i,k}, R_{j,k}|\gamma_k \sim \begin{cases} \mathcal{U}_{[-\delta,\delta]} & \text{si } \gamma_k = 1 \\ \mathcal{U}_{[-\kappa,\kappa]} & \text{si } \gamma_k = 0 \end{cases}$$

La loi a priori sur le vecteur γ est elle aussi simple :

$$\gamma = (\underbrace{1,\ldots,1}_{k \text{ fois}},\underbrace{0,\ldots,0}_{\mathfrak{m}-k \text{ fois}}) \text{ avec probabilité } \frac{\beta^{k-1}(1-\beta)}{1-\beta^{\mathfrak{m}}}.$$

On a enfin besoin d'une hypothèse sur le bruit, qui doit être sous-exponentiel et qui est donc similaire à l'hypothèse nécessaire au théorème 1.2.

On peut maintenant énoncer le théorème en introduisant l'ensemble $\mathcal{M}(k)$, sous-ensemble de matrices de rang au plus k (α est un réel positif) :

$$\mathcal{M}(k) = \left\{LR^\top, (L,R) \in \mathbb{R}^{m_1\times k} \times \mathbb{R}^{m_2\times k}, (\|L\|_\infty, \|R\|_\infty) \leq \alpha/\mathfrak{m}\right\}.$$

Théorème 1.5 (Théorème 1 de [Mai and Alquier, 2015])**.** *Supposons qu'on observe $(X_i, Y_i)_{1\leq i\leq n}$ suivant (1.11) et que l'hypothèse 1.3 soit satisfaite. Supposons que $\|M^\star\|_\infty \leq \boldsymbol{a}$. Prenons $\tau = n/2C$, δ, κ étant aussi fixés. Alors, pour tout $\epsilon \in]0,1[$, dès que $n > \mathfrak{M}$, on a avec probabilité supérieure à $1-\epsilon$:*

$$\left\|\widehat{M}_\tau - M^\star\right\|_{S_2,\Pi} \leq \min_{k\in[\mathfrak{m}]} \left\{ 3 \inf_{M\in\mathcal{M}(k)} \|M - M^\star\|_{S_2,\Pi} + C_2 \frac{(m_1+m_2)k\log\mathfrak{m}}{n} + \frac{C_3 \log\frac{2}{\epsilon}}{n} \right\},$$

où C, C_2, C_3 sont des constantes connues dépendant de L, β et du bruit.

Contrairement aux résultats fréquentistes, c'est une inégalité oracle dans le sens où elle ne demande pas que la matrice $M^\star$ soit de faible rang. Ainsi, si la matrice est de plein rang mais peut être bien approchée par une matrice de faible rang (typiquement le cas quand les valeurs singulières décroissent rapidement), alors la vitesse de convergence est la meilleure possible en faisant l'arbitrage entre biais et variance. Cette vitesse est optimale à un logarithme près. De plus, les conditions sur le tirage sont minimales et permettent facilement de faire les hypothèses nécessaires pour une reconstruction suivant la norme de Frobenius. La question du calcul de l'estimateur reste en revanche ouverte. La première méthode expliquée plus loin, le MCMC, a été mise en œuvre dans l'article et cela fonctionne bien sur des simulations. En revanche, c'est une méthode relativement lente pour un paramètre de grande taille. Nous verrons alors une seconde méthode approchée plus rapide.

1.3.2 Calcul de l'estimateur bayésien : MCMC

La méthode la plus courante à l'heure actuelle pour calculer une valeur approchée de l'estimateur bayésien est d'utiliser une méthode de MCMC (Markov Chain Monte Carlo). L'idée est d'effectuer un grand nombre de tirage suivant une chaîne de Markov dont la loi invariante est la loi a posteriori. Si on note $(\theta^t)_{1\leq t\leq T}$ l'échantillon de taille T tiré, l'estimateur bayésien sera approché par :

$$\tilde{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \theta^t.$$

Si la chaîne construite a les bonnes propriétés, le théorème ergodique permet d'affirmer que $\tilde{\theta}_T$ tend vers $\widehat{\theta}_\tau$ quand T tend vers $+\infty$. Le lecteur

intéressé pourra se référer à [Robert and Casella, 2005] pour une longue et complète introduction aux méthodes MCMC.

Parmi les nombreuses méthodes qui existent, le modèle de complétion de matrice tel que présenté ici est assez facile à traiter car le bruit est gaussien et la loi a priori sur les facteurs L et R aussi. Les lois conditionnelles s'expriment donc facilement (en choisissant π^γ judicieusement également) et il est assez facile de construire un Gibbs Sampler (ou échantillonneur de Gibbs en français) : cela est fait, avec quelques adaptations mineures, dans l'article [Salakhutdinov and Mnih, 2008]. Le modèle est appliqué aux données du prix Netflix et a de bonnes performances en prédictions. Néanmoins, vu la taille des données et de la matrice à compléter (donc le paramètre), cette méthode qui est très intensive en calcul, touche ses limites. De plus, les garanties théoriques concernant la méthode d'approximation sont très complexes à obtenir. On présente donc maintenant une méthode de calcul approchée qui est beaucoup plus rapide.

1.3.3 Calcul de l'estimateur : Approximation variationnelle

Il est courant en mathématiques, quand une quantité est difficile à calculer, de chercher une approximation. Cette idée est à la base des approximations variationnelles bayésiennes (dénommées plus tard en *VB*) où la loi a posteriori $\widehat{\rho}_\tau$ est approchée par une fonction dont il est plus facile de calculer l'espérance (pour une bonne entrée en matière, le lecteur est invité à se référer à [Bishop, 2006], chapitre 10 ; la méthode est populaire au point d'avoir été présentée dans un *tutorial* à NIPS en 2016). Pour définir une approximation, il faut une mesure de dissimilarité entre les objets et une famille d'approximation. Pour la dissimilarité, on prendra ici la divergence de Kullback-Leibler bien que d'autres fonctions soient possibles. Si μ est une mesure dominée par ν, on définit alors la divergence de Kullback-Leibler $\mathcal{K}(\mu, \nu)$ par :

$$\mathcal{K}(\mu, \nu) = \int \log \frac{d\mu}{d\nu} d\mu.$$

Étant donné une famille d'approximation $\mathcal{F}$, on cherche alors :

$$\widetilde{\rho}_\tau = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \widehat{\rho}_\tau). \tag{1.12}$$

L'estimateur approché (qui dépend donc de la famille choisie $\mathcal{F}$) sera

alors la moyenne par rapport à la loi approché :

$$\widetilde{\theta}_\tau = \int \theta \widetilde{\rho}_\tau(d\theta).$$

La famille $\mathcal{F}$ doit être petite pour que le minimiseur soit facile à calculer (et accessoirement l'espérance par rapport à cette mesure) mais en même temps doit être suffisamment grande pour que l'approximation soit de bonne qualité. L'intérêt d'une telle méthode est qu'elle est habituellement beaucoup plus rapide que les méthodes MCMC. Inversement, si la précision de l'estimation par MCMC peut s'améliorer en augmentant la taille de l'échantillon, l'approximation variationnelle dépend de la famille $\mathcal{F}$ et il peut être difficile de l'améliorer tout en gardant un programme faisable. Dans la suite, on ne traitera que le cas où les mesures ont une densité par rapport à une mesure de référence.

Une première famille populaire dans la littérature est la famille des lois indépendantes (cette méthodologie est appelée *Mean Field* dans la littérature). On se donne un découpage de θ en blocs $(\theta_i)_{i\in[N]}$ et on pose alors :

$$\mathcal{F}^{MF} = \left\{ q : q(\theta) = \prod_{i=1}^{N} q_i(\theta_i) \right\}.$$

Intuitivement, la distribution jointe $\widehat{\rho}_\tau$ est approchée par une loi ayant des composantes indépendantes. Le graphique 1.1 montre l'approximation d'une loi normale bivariée (aux composantes volontairement corrélées) par une loi normale dont les composantes sont indépendantes. De part la forme factorisée des éléments de $\mathcal{F}$, il vient alors que $\widetilde{q}$ est un point fixe satisfaisant :

$$\forall i \in [N],\ q_i(\theta_i) \propto \exp\left(\int \{-\tau r(\theta) + \log \pi(\theta)\} \prod_{j\neq i} q_j(\theta_j) d\theta_j \right).$$

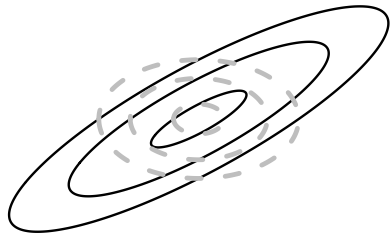
Si la loi a priori est choisie de façon appropriée, les mesures q_j sont alors paramétriques et l'optimisation revient à chercher un point fixe sur un paramètre de dimension finie.

Une autre possibilité est une famille paramétrique. On pose alors :

$$\mathcal{F}^{P} = \{ f_m : m \in \mathcal{M} \},$$

où $\mathcal{M}$ est de dimension finie. Pour calculer pratiquement l'élément optimal, on va utiliser l'identité suivante :

$$\widetilde{\rho}_\tau = \arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \widehat{\rho}_\tau) = \arg\min_{\rho\in\mathcal{F}} \left\{ \int \tau r(\theta) \rho(d\theta) + \mathcal{K}(\rho, \pi) \right\}. \quad (1.13)$$



Graphique 1.1 – Exemple d'approximation variationnelle d'une loi normale ayant un coefficient de corrélation de 0.85 par des composantes indépendantes.

Le membre de droite de (1.13) peut être alors facile à calculer, ou alors être lui-même borné par une quantité qui sera aisément optimisable. Le lecteur est renvoyé à l'article [Alquier et al., 2016] pour une explication détaillée. Cet article donne en plus une méthode pour calculer des bornes sur le risque de l'estimateur obtenu. Des idées similaires seront utilisées dans le chapitre 3.

Le modèle de complétion de matrice avec la perte quadratique a été traité par cette méthode sous une approche de type *mean-field* par Lim and Teh [2007]. Comme toutes les lois sont bien choisies, les calculs sont explicites et l'algorithme est alors directement construit. Il s'avère qu'il converge rapidement même pour un grand jeu de données. Le problème ici est qu'on n'a pas de garanties de convergence de l'algorithme ni de garanties sur l'estimateur obtenu : il est en effet différent de l'estimateur bayésien classique. Dans le chapitre 3, nous développons un estimateur variationnel pour la complétion de matrice binaire et nous obtenons ses propriétés théoriques.

1.4 Résumé (substantiel) des chapitres

1.4.1 Chapitre 3 : *1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation*

Ce chapitre propose un estimateur pour la complétion de matrice binaire (où les entrées observées sont dans $\{-1, +1\}$. Les travaux précédents à l'aide d'un modèle de régression généralisé de type *Logit* ne sont pas complètement concluants car ils s'intéressent à la reconstruction de la matrice de paramètre et non au risque en prédiction. Les observations étant n couples i.i.d $(X_i, Y_i) \in \mathcal{X} \times \{-1, +1\}$, le but est de trouver un estimateur dont le risque 0/1 s'approche du risque minimum, qui est minimisé par M^B (le prédicteur de Bayes) : $M^B_{i,j} = \operatorname{sign} \mathbb{E}(Y|X = E_{i,j})$.

Estimateur et résultats théoriques.

La première étape est la construction de l'estimateur. On commence par définir grâce à la perte charnière (plus connue sous son appelation anglo-saxonne de *hinge loss*), le risque empirique pour une matrice $M \in \mathbb{R}^{m_1 \times m_2}$:

$$\forall M \in \mathbb{R}^{m_1 \times m_2}, \quad r^h(M) = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - Y_i \langle M, X_i \rangle).$$

Nous utilisons la loi a priori définie par (1.10) comme loi favorisant les matrices de faible rang avec l'ajout d'un paramètre. Le paramètre final est donc $\theta = (L, R, \gamma)$. La loi pseudo a posteriori est alors définie par (1.9). Comme il est extrêmement coûteux pour ce problème d'utiliser une méthode MCMC pour approcher la loi, nous utilisons une approximation variationnelle similaire à (1.12) pour une famille paramétrique $\mathcal{F}^P$ finement choisie. La borne

$$\forall f_m \in \mathcal{F}^P, \int \tau r^h(\theta) f_m(d\theta) + \mathcal{K}(f_m, \pi) = \mathcal{L}(m)$$

n'est pas calculable explicitement en fonction des paramètres m car la matrice M est factorisée. Nous utilisons ici une majoration de la borne

$$\mathcal{L}(m) \leq AVB(m)$$

qui a, elle, une forme explicite. La loi qui sera utilisée finalement est la loi dont le paramètre minimise cette borne $f_{\widehat{m}}$: $\widehat{m} = \min_{m \in \mathcal{M}} AVB(m)$

et l'estimateur est défini par :

$$\widehat{M} = \int LR^\top f_{\widehat{m}}(d\theta).$$

Nous pouvons travailler sur les propriétés théoriques de cet estimateur, dans l'esprit de l'article [Alquier et al., 2016]. Nous trouvons une borne empirique, dans le sens où elle est calculable sur les données, et une borne théorique. Celle-ci nécessite une hypothèse de marge, ce qui est classique pour les problèmes de classification. Ces deux bornes portent sur le risque de classification $R_{0/1}$ définie par (1.8). On rappelle ici le théorème 2, qui est la borne théorique. On note $\overline{R} = \inf_M R_{0/1}(M)$.

Théorème 1.6 (Théorème 2 du chapitre 3)**.** *Supposons que l'hypothèse de marge soit satisfaite pour un $C > 0$. Alors, pour tout $\varepsilon, s \in]0,1[$, pour $\lambda = sn/C$, on a l'inégalité suivante avec probabilité au moins $1-\varepsilon$:*

$$\int R_{0/1} df_{\widehat{m}} \leq 2(1+3s)\overline{R} + \mathcal{C}\left(\frac{\operatorname{rang}(M^B)\,\mathfrak{M}(\log n + L(M^B)) + \log 1/\varepsilon}{n}\right),$$

où L est une fonction déterministe et $\mathcal{C}$ est une constante dépendant des constantes connues et de la loi a priori sur γ.

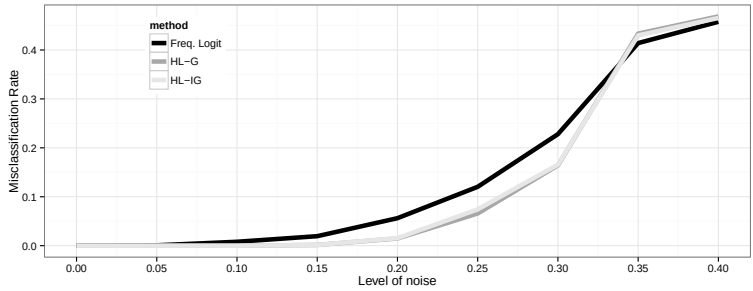
Par rapport aux résultats précédents comme celui du théorème 1.4, ce résultat permet de comparer le risque 0/1 intégré de l'estimateur au meilleur risque possible. La limite de ce résultat est que, théoriquement, le facteur 2 ne montre pas la consistance de notre estimateur en toute géénralité. Un cas particulier régulièrement étudié est le cas sans bruit. Dans ce cas, l'estimateur est consistant et la vitesse est minimax à un terme logarithmique près.

Utilisation en pratique.

La quantité $AVB(m)$ n'est pas convexe en m car la matrice M est factorisée en un produit $LR^\top$. Néanmoins, elle est biconvexe dans le sens où elle est convexe en L et en R. Nous proposons alors un algorithme d'optimisation coordonnées par coordonnées où, pour les coordonnées où les minimums ne sont pas explicites, l'étape est remplacée par une descente de sous-gradient.

Cela nous permet alors de tester l'estimateur sur des données simulées dans un premier temps. Plusieurs scénarios sont testés : observations suivant le modèle logistique ou non, matrice de Bayes de faible rang ou

non. En sortant du modèle logistique, on voit que l'estimateur proposé fait mieux que celui provenant de la régression généralisée. Par exemple, sur le graphique 1.2, pour la reconstruction d'une matrice de rang 3, les performances sont bien meilleures pour des valeurs médianes. Les différents algorithmes sont enfin testés sur la base de données MovieLens et montrent leurs capacités à être appliqués sur de grandes bases de données.



Graphique 1.2 – Risque 0/1 pour la reconstruction d'une matrice de rang 3 (taille 200 par 200) en fonction de l'ampleur du bruit *switch* (avec probabilité p la valeur observée est l'opposé de la vraie valeur).

Approximation variationnelle pour le modèle logistique.

Ce chapitre est aussi l'occasion de développer l'approximation variationnelle pour le modèle logistique. Le modèle de complétion usuel avec un bruit gaussien a vu son approximation variationnelle développée dans [Lim and Teh, 2007]. Ici, il faut utiliser une double approximation. Nous pouvons aussi proposer des lois a priori différentes pour γ. Les approximations variationnelles sont alors trouvées en utilisant la famille de loi *inverse gaussienne généralisée*.

1.4.2 Chapitre 4 : *Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions*

Ce chapitre apporte un autre éclairage sur le problème de complétion de matrice binaire, mais permet aussi de traiter d'autres problèmes. Il s'intéresse en effet aux propriétés du minimiseur du risque empirique

régularisé quand la fonction de perte utilisée est lipschitzienne. Si on note $\ell(f(x), y)$ la perte pour une prédiction par f en (x, y), cette propriété s'exprime par :

$$\forall (x,y), \forall (f_1, f_2), \quad |\ell(f_1(x), y) - \ell(f_2(x), y)| \leq |f_1(x) - f_2(x)| \,.$$

Ce caractère lipschitzien s'applique par exemple aux fonctions de perte suivantes :

- perte pour la régression quantile de paramètre $\tau \in (0,1)$: $\ell(y', y) = \tau \max(0, y - y') + (1 - \tau) \max(0, y' - y)$ pour $y', y \in \mathbb{R}$
- perte logistique : $\ell(y', y) = \log(1 + \exp(-yy'))$ pour $(y, y') \in \{-1, +1\} \times \mathbb{R}$
- perte charnière : $\ell(y', y) = \max(0, 1 - yy')$ pour $(y, y') \in \{-1, +1\} \times \mathbb{R}$

On note $(X_i, Y_i)_{i=1}^n$ les observations. En munissant l'espace des prédicteurs $\mathcal{F}$ d'une norme, l'estimateur que l'on considèrera est :

$$\widehat{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \|f\| \right\}.$$

Si le minimiseur du risque empirique régularisé a été beaucoup étudié pour la perte des moindres carrés, les analyses sont plus réduites pour une fonction de perte lipschitzienne. Pourtant, cette caractéristique est très courante dans le cadre de la classification (pour $y_i \in \{-1, +1\}$) avec les fonctions de perte logistique et charnière. La régression quantile recouvre la perte ℓ_1, qui est une version robuste des moindres carrés, mais permet aussi de traiter n'importe quel quantile entre 0 et 1.

Résultats théoriques.

Les résultats théoriques sont développés dans deux cadres en fonction des hypothèses à faire sur X et $\mathcal{F}$: le cadre sous-gaussien et le cadre borné. Comme les résultats sont généraux, ils dépendent de quantités dépendant essentiellement de $(\mathcal{F}, \|\cdot\|)$. On note l'oracle $f^\star$ et l'excès de risque de $f \in \mathcal{F}$ par $\mathcal{E}(f)$ où :

$$f^\star = \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)], \quad \mathcal{E}(f) = \mathbb{E}[\ell(f(X), Y)] - \mathbb{E}[\ell(f^\star(X), Y)].$$

Nous esquissons ici le résumé de la stratégie à adopter pour calculer effectivement les bornes non asymptotiques. Les définitions formelles des quantités se trouvent dans le chapitre.

1. Trouver les paramètres de Bernstein (κ, A) associé à la perte ℓ et la famille $\mathcal{F}$.
2. Calculer la fonction de complexité

$$r(\rho) = \left[\frac{A\rho \mathrm{comp}(B)}{\sqrt{n}}\right]^{1/2\kappa},$$

où $\mathrm{comp}(B)$ dépend du cadre (sous-gaussien ou borné) et de la boule unité $B = \{f : \|f\| \leq 1\}$.

3. Après le calcul du sous-différentiel de la norme $\partial \|\cdot\| (f^\star)$, il faut résoudre l'équation de sparsité. C'est à cet endroit que la parcimonie de l'oracle peut jouer un rôle. $\rho^\star$ est alors le rayon satisfaisant :

$$\Delta(\rho^\star) = 4\rho^\star/5.$$

4. On peut maintenant appliquer les théorèmes en fonction du cadre. On alors, avec grande probabilité :

$$\left\|\widehat{f} - f^\star\right\| \leq \rho^\star, \quad (\mathbb{E}[\widehat{f}(X) - f^\star(X)]^2)^{1/2} \leq r(2\rho^\star),$$
$$\mathcal{E}(\widehat{f}) \leq C[r(2\rho^\star)]^{2\kappa}.$$

Ces résultats sont obtenus sans hypothèse sur Y, ce qui montre la robustesse des fonctions de perte lipschitziennes. Nous pouvons alors appliquer ces résultats à deux cadres, la régression logistique et la complétion de matrice, en calculant les quantités nécessaires. Dans le chapitre, un autre exemple sur les RKHS est aussi développé.

Application à la régression logistique.

Les résultats connus de régression logistique pénalisée utilisent généralement la norme ℓ_1 sur les vecteurs. La norme SLOPE, introduite par [Bogdan et al., 2015], permet d'atteindre une vitesse minimax pour la régression aux moindres carrés pénalisés mais n'avait encore jamais été étudiée avec la perte logistique. On la définit, pour $t \in \mathbb{R}^p$ par

$$\|t\|_{SLOPE} = \sum_{j=1}^{p} \sqrt{\log(ep/j)} t_{(j)},$$

où les $t_{(j)}$ sont un arrangement décroissant des $(|t_j|)$. Les observations (X_i, Y_i) sont à valeur dans $\{-1, +1\} \times \mathbb{R}^p$ et l'estimateur considéré est

alors :

$$\widehat{t} = \underset{t\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-Y_i \langle t, X_i\rangle) + \frac{c}{\sqrt{n}} \|t\|_{SLOPE} \right\}.$$

Dans le cadre sous-gaussien, qu'on considère ici, comp(B) est de l'ordre d'une constante. La constante de Bernstein κ vaut 1. Comme la norme SLOPE a un grand sous-différentiel quand des coordonnées sont nulles, la solution de l'équation de sparsité donne $\rho^\star = c\frac{s}{\sqrt{n}} \log \frac{ep}{s}$ où s est le nombre de coordonnées non nulles de $t^\star$. Finalement, avec grande probabilité, on a :

$$\mathcal{E}_{\text{logistic}}(\widehat{t}) \leq c\frac{s\log ep/s}{n}$$

Ce résultat améliore les résultats connus utilisant la norme ℓ_1 : l'excès de risque était contrôlé par une borne de l'ordre $s\log p/n$. Quand s est plutôt grand, de l'ordre d'une fraction de p, le résultat obtenu avec la norme SLOPE est meilleur. Par ailleurs, le même genre de résultats peut être calculé pour la perte régression quantile.

Application à la complétion de matrice.

Cette partie illustre le cadre borné, et permet d'avoir des résultats nouveaux pour la complétion de matrice. Les prédicteurs sont les matrices M de taille $m_1 \times m_2$ où les entrées sont bornées par $\mathbf{a}$. Les variables X_i sont à valeur dans $\mathcal{X}$, l'ensemble des matrices masque. La norme de régularisation utilisée est la norme nucléaire. Les résultats ici ne sont écrits que pour $\kappa = 1$ soit la vitesse rapide, les résultats généraux sont dans le chapitre. On considère alors l'estimateur suivant :

$$\widehat{M} = \underset{M:\|M\|_\infty \leq \mathbf{a}}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(\langle M, X_i\rangle, Y_i) + \lambda \|M\|_{S_1} \right\}. \tag{1.14}$$

Le paramètre de complexité est :

$$r(\rho) = c\left[\rho\sqrt{\frac{\log(m_1+m_2)}{n\,\mathfrak{m}}}\right]^{1/2}.$$

La norme nucléaire a un sous-différentiel grand quand la matrice est de faible rang. Si la matrice oracle est de rang s, l'équation de sparsité

est vérifiée par :

$$\rho^\star = csm_1m_2\left(\frac{\log(m_1+m_2)}{n\,\mathfrak{m}}\right)^{\frac{1}{2}}$$

Finalement, avec grande probabilité, $\widehat{M}$ satisfaisant (1.14) est telle que :

$$\frac{1}{m_1m_2}\left\|\widehat{M}-M^\star\right\|_{S_2}^2 \leq c\frac{s\,\mathfrak{M}\log(m_1+m_2)}{n}$$
$$\mathcal{E}(\widehat{M}) \leq c\frac{s\,\mathfrak{M}\log(m_1+m_2)}{n}$$

Ce résultat est intéressant car il permet non seulement d'avoir un résultat sur la reconstruction de $M^\star$ tel que souvent recherché dans la complétion de matrice, mais aussi de contrôler l'excès de risque. Il est possible d'appliquer ce résultat à plusieurs fonctions de pertes.

Complétion de matrice binaire.

En complétion de matrice binaire (quand Y_i est à valeur dans $\{-1,+1\}$), on peut utiliser la perte logistique ou la perte charnière. Avec la perte logistique, κ vaut 1 et on retrouve alors le résultat de [Klopp et al., 2015], mais on obtient aussi une borne sur le risque logistique de $\widehat{M}$. À l'aide du résultat de [Zhang, 2004], le risque 0/1 de $\widehat{M}$ est alors majoré avec grande probabilité par $\sqrt{s\,\mathfrak{M}\log(m_1+m_2)/n}$.

Il est alors utile de considérer la perte charnière. Le paramètre de Bernstein vaut 1 en faisant l'hypothèse que $M^\star$ n'a pas d'entrées trop proches de 0. Avec cette hypothèse, l'estimateur (1.14) a comme propriété qu'avec grande probabilité, l'excès de risque charnière est borné par $s\,\mathfrak{M}\log(m_1+m_2)/n$. L'excès de risque 0/1 est alors contrôlé par la même quantité et est bien meilleur que tous les résultats connus.

Le chapitre présente aussi des bornes inférieures ainsi qu'un ensemble de simulations. Nous proposons un algorithme de type *Alternating Direction Method of Multipliers* pour calculer $\widehat{M}$ qui est proche de l'algorithme 1.1. Nous le testons sur les données de MovieLens. Les notebooks se trouvent à l'adresse http ://sites.google.com/site/vincentcottet/code.

Complétion de matrice à l'aide de la perte régression quantile.

Cette fonction de perte est rarement utilisé en complétion de matrice alors qu'elle peut avoir plusieurs utilités. Pour $\tau = 1/2$, la fonction de

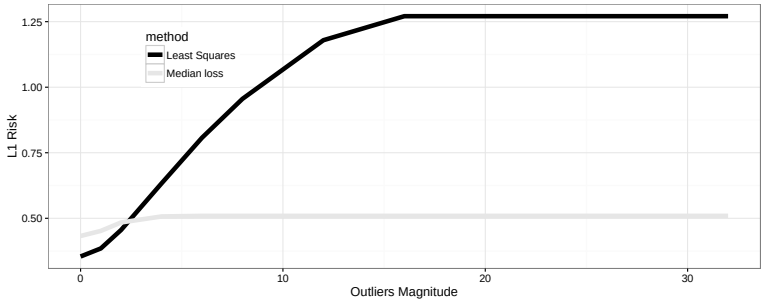
perte correspond à la valeur absolue : c'est une fonction de perte qui est beaucoup plus robuste aux valeurs aberrantes que la perte quadratique (de la même façon que la médiane n'est pas impactée alors que la moyenne l'est). Pour d'autres valeurs de τ, cela permet de contrôler des quantiles de la même façon que les régressions quantiles.

Le paramètre de Bernstein κ vaut 1 dès que le bruit a une densité par rapporte à la mesure de Lebesgue qui est bornée inférieurement sur un domaine suffisament large (cela est facilement calculable pour des lois telles que normale, Student et même Cauchy). Les résultats sont alors immédiats en appliquant le théorème principal. Pour $\tau = 1/2$, si le bruit est centré, la matrice à reconstruire est la même que pour la perte quadratique. La reconstruction en norme de Frobenius est immédiat et la vitesse est la même que celle de [Klopp, 2014] ou de [Mai and Alquier, 2015] et est optimale à un terme logarithmique près. L'avantage est que les hypothèses sur le bruit sont beaucoup plus larges (le bruit pour la perte quadratique doit être sous-exponentiel).

Nous testons alors ce caractère robuste grâce à plusieurs scénarios. L'un d'entre eux concerne la présence de points aberrants. Nous partons d'une matrice de rang 3 de taille 200×200 et augmentons la proportion de points aberrants. Les performance de l'estimateur utilisant les moindres carrés se dégradent rapidement et pour une proportion supérieure à 10%, la matrice estimée est nulle partout. Inversement, l'estimateur de la médiane ($\tau = 1/2$) voit ses performances baisser très légèrement au début puis rester stable jusqu'à un niveau massif de points aberrants (30%). Dans le même ordre d'idée, l'ampleur des points aberrants n'a que peu d'impact sur l'estimateur de la médiane comme on peut le voir sur le graphique 1.3.

1.4.3 Chapitre 5 : *Divide and Conquer in ABC : Expectation-Propagation algorithms for likelihood-free inference*

Les deux chapitres précédents traitent des problèmes de grande dimension : le paramètre est très grand et on en cherche un modèle parcimonieux qui colle aux données. Les méthodes ABC, *Approximate Bayesian Computation*, se sont développées grâce à la puissance informatique maintenant disponible et permettent d'aborder des problèmes nouveaux. Elles visent à faire de l'inférence bayésienne pour des modèles dont la vraisemblance n'est pas explicite ou qui est trop coûteuse à calculer. Cela provient souvent de la constante de normalisation de la vraisemblance, qui dépend de θ, est inaccessible. Ces modèles apparaissent fréquemment



Graphique 1.3 – Reconstruction en norme ℓ_1 d'une matrice de faible rang. L'ampleur des 10 % des valeurs aberrantes augmente de 0 à 30.

en biologie, en neuroscience ou dans des modèles spatiaux. Tout ce qui est requis pour l'algorithme ABC est de pouvoir simuler suivant le modèle : l'évaluation de la vraisemblance n'est pas nécessaire.

Le chapitre présenté ici est une extension de [Barthelmé and Chopin, 2014]. Nous allons commencer par présenter la méthode ABC puis nous expliquerons l'approximation EP pour aboutir à la description de l'algorithme finalement proposé. Nous exposerons enfin l'exemple illustratif.

L'algorithme ABC.

L'algorithme ABC de base est très simple et paraît tout à fait naturel. On se donne une loi a priori $p(\theta)$, une vraisemblance $p(x|\theta)$ et des observations $x^\star$ provenant de ce modèle. Il faut alors se fixer une distance entre les données observées et simulées $d(x^\star, x)$ et un seuil $\varepsilon > 0$. L'algorithme est de répéter les étapes suivantes tant qu'on n'a pas assez de valeurs acceptées :

1. tirer $\theta \sim p(\theta)$
2. tirer des données suivant la vraisemblance $x \sim p(x|\theta)$
3. accepter θ si $d(x^\star, x) \leq \varepsilon$.

Finalement, les valeurs acceptées constituent un échantillon de la loi :

$$p_\varepsilon(\theta|x^\star) \propto p(\theta) \int p(x|\theta) \mathbb{1}_{\{d(x,x^\star)\leq\varepsilon\}} dx.$$

Une limitation importante de cette méthode est de devoir choisir une distance d. S'il est possible de prendre $d(x^\star, x) = \|s(x^\star) - s(x)\|$ où s

est une statistique exhaustive, alors l'approximation tendra vers la loi a posteriori si ε tend vers 0. Si ce n'est pas le cas, l'approximation sera double et il sera très difficile de la contrôler.

Diviser pour régner : découper le problème en blocs.

Pourtant, dans beaucoup de cas, les données forment naturellement des blocs qui seront notés $(x_i)_{i=1}^k$; chaque bloc peut être de taille différente et on note $x_{[k]} = (x_1, \ldots, x_k)$. La vraisemblance peut s'écrire de manière séquentielle par bloc [6] :

$$p(x|\theta) = \prod_{i=1}^{k} p(x_i|x_{[i-1]}, \theta).$$

S'il est possible de simuler x_i suivant $p(x_i|x_{[i-1]}, \theta)$, on peut alors découper le problème en plusieurs blocs et approcher la loi :

$$p_\varepsilon(\theta|x^\star) \propto p(\theta) \prod_{i=1}^{k} \int p(x_i|x^\star_{[i-1]}, \theta) \mathbb{1}_{\{\|s_i(x_i) - s_i(x_i^\star)\| \leq \varepsilon\}} dx_i, \qquad (1.15)$$

où s_i est une statistique pour x_i. L'algorithme EP-ABC propose de faire une approximation de type EP de la loi $p_\varepsilon(\theta|x^\star)$.

Le premier cas où la vraisemblance apparaît naturellement de manière factorisée est quand les données sont i.i.d. ou i.i.d. par blocs. Par exemple, si on mesure plusieurs années de suite un phénomène géographique comme dans l'application présentée plus loin, on pourra faire l'hypothèse que les observations de chaque année sont indépendantes. Un deuxième cas très fréquent se trouve quand les données sont une ou des séries temporelles.

***Expectation Propagation* : une approximation bayésienne.**

L'algorithme EP, *Expectation Propagation*, est assez similaire dans l'esprit à l'approche VB : le but est de trouver une loi q dans une certaine famille paramétrique $\mathcal{Q}$ qui sera proche de la loi cible π. Il est utilisé quand π se factorise en n facteurs :

$$\pi(\theta) = \frac{1}{Z_\pi} \prod_{i=0}^{k} l_i(\theta).$$

6. Cette écriture est toujours correcte et ne requiert aucune hypothèse.

L'idée est alors d'approximer chaque *site* l_i par un facteur q_i et l'approximation finale sera

$$q(\theta) = \prod_{i=0}^{k} q_i(\theta).$$

Afin de calculer cela, l'algorithme ressemble à une optimisation coordonnées par coordonnées en effectuant plusieurs passages sur les données tant qu'on n'a pas convergé. La mise à jour du site i se fait par les étapes suivantes :

1. calculer la *cavity distribution* $q_{-i}(\theta) \propto \prod_{j \neq i} q_j(\theta)$. La loi hybride sera $h_i \propto q_{-i}(\theta) l_i(\theta)$.
2. mettre à jour q_i tel que

$$q \in \arg\min_{f \in \mathcal{Q}} \mathcal{K}(h_i, f). \tag{1.16}$$

La dernière étape est cruciale et a une forme explicite si $\mathcal{Q}$ est une famille exponentielle. Dans ce cas, cela revient à faire correspondre les moments entre h_i et q. Le chapitre utilise des lois normales comme approximation ; l'étape principale (1.16) revient alors à calculer les moments :

$$\mu_h = \frac{1}{\int h_i(\theta) d\theta} \int \theta h_i(\theta) d\theta$$
$$\Sigma_h = \frac{1}{\int h_i(\theta) d\theta} \int \theta\theta^\top h_i(\theta) d\theta - \mu_h \mu_h^\top$$

En pratique, les résultats sont plutôt bons et l'approximation se révèle souvent être meilleure qu'une approximation VB. En effet, l'algorithme EP conserve une covariance pleine pour le paramètre, et donc entre ses coordonnées. À l'inverse, l'approximation VB de type *mean-field* ne va pas conserver cette souplesse et va l'approcher par des coordonnées indépendantes. Néanmoins, dans une approche de prédiction, il n'est parfois pas nécessaire d'avoir cette information et donc l'approche VB peut se révéler pertinente.

L'algorithme EP souffre pourtant d'un manque d'études théoriques. Deux articles récents [Dehaene and Barthelmé, 2015b] et [Dehaene and Barthelmé, 2015a] effectuent une avancée importante : le premier vise à étudier la qualité de l'approximation d'un point fixe de l'algorithme. Les auteurs montrent que la moyenne et la variance, dans un cas d'approximation gaussienne, sont proches de celles de la vraie loi. Le second

s'intéresse à l'algorithme et montrent qu'il peut s'interpréter comme un algorithme de Newton d'optimisation. Cela peut entraîner une certaine instabilité des mises à jour et une difficulté à converger qui peut être réparée en moyennant les mises à jour.

L'algorithme EP-ABC en pratique.

La loi cible définie en (1.15) est factorisée en $k+1$ blocs (le premier étant la loi a priori). Chaque site est :

$$l_i(\theta) = \int p(x_i|x^\star_{[i-1]}, \theta) \mathbb{1}_{\{\|s_i(x_i) - s_i(x_i^\star)\| \leq \varepsilon\}} dx_i.$$

La faisabilité de l'approximation EP revient au calcul des moments des lois hybrides. Pour cela, nous utilisons une approximation de type ABC, qui est possible si la simulation de données suivant la densité $p(x_i|x^\star_{[i-1]})$ est faisable. L'algorithme complet se trouve dans le chapitre.

Les avantages de cet algorithme sont importants : à chaque étape des θ sont tirés suivant les lois cavité q_i qui sont normalement plus précises que la loi a priori [7]. Ensuite, la simulation de x_i et l'acceptation de θ associé se fait en fonction de $\|s_i(x_i) - s_i(x_i^\star)\|$. Comme on réduit la dimension, la probabilité d'accepter est plus grande, ou alors il est possible de réduire la tolérance. Cela permet de ne pas souffrir du fléau de la dimension : pour l'ABC, l'ajout de points entraîne habituellement une diminution du taux d'acceptation et une détérioration de l'approximation [8]. Avec l'algorithme EP-ABC, le découpage en blocs entraîne la neutralité de l'ajout de nouveaux points s'ils forment de nouveaux blocs.

Enfin, cet algorithme est facilement parallélisable, au prix d'une petite adaptation. L'étape la plus coûteuse, qui est le calcul des moments, peut se faire en parallèle grâce à la segmentation du problème en blocs. De plus, le parallélisme lisse les mises à jour et stabilise l'algorithme.

Une application sur des extrêmes spatiaux.

Dans l'article [Barthelmé and Chopin, 2014] sont traités trois exemples : n observations i.i.d. provenant d'une loi alpha-stable (distribution dont la densité n'a pas de forme explicite), un modèle Lotka-Volterra de proie-prédateurs et un modèle de temps de réactions à un stimulus où sont

7. Ce point est souvent amélioré dans des versions plus évoluées de l'algorithme ABC.

8. Ici la dimension concerne le nombre de points et non la taille du paramètre.

modélisés des séries temporelles évoluant conjointement. On voit dans ces deux derniers exemples le caractère séquentiel qui apparaît naturellement.

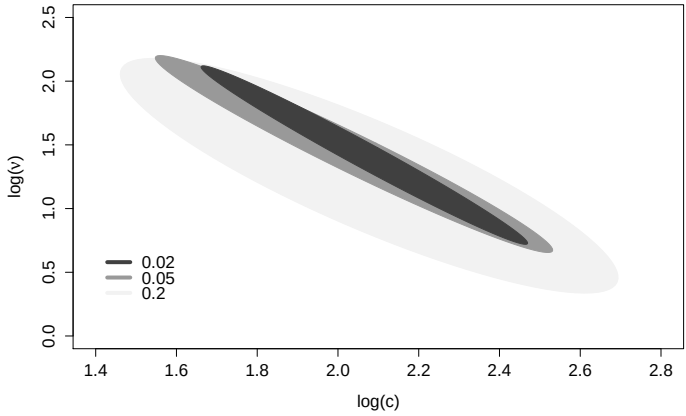
Dans le chapitre 5, nous proposons un exemple où cette fois-ci les données sont k blocs indépendants de n points qui sont interdépendants [9]. Ici, les observations sont des réalisations d'un processus spatial qui est observé en différents points de l'espace notés $(x_i)_{i=1}^n$. L'observation pour l'année j au points x_i est $y_j(x_i)$ et on fait l'hypothèse que les observations sont indépendantes suivant les années. Le processus considéré ici est un processus max-stable ; ces processus sont utilisés pour modéliser des extremums. Le but ici est d'inférer la covariance du processus pour avoir une idée de la dépendance entre les points. Pour plus de deux points la vraisemblance n'est pas explicite. Il a été proposé une méthode de vraisemblance composite, qui est une méthode fréquentiste approchée. Cette méthode est asymptotiquement valable mais nous ne nous trouvons pas vraiment dans ce cadre ici.

La méthode ABC classique a été proposée pour ce modèle dans [Erhardt and Smith, 2012]. Elle est non seulement extrêmement coûteuse car la génération de processus max-stable est très chère mais souffre aussi d'un taux d'acceptation très faible. Néanmoins, cet article propose plusieurs statistiques et est un bon point de départ pour appliquer notre méthode.

L'algorithme EP-ABC s'applique ici très bien étant donné le caractère indépendant des blocs. Cela permet de les traiter indépendamment et d'avoir une étape d'approximation en moins. Il est également possible ici d'utiliser le recyclage des simulations, ce qui est un autre avantage de cette méthode. Finalement, nous appliquons ce modèle sur des données de précipitations maximales en Suisse. Le maximum est calculé entre juin et août en 79 points entre 1962 et 2008. On choisit la fonction de variance de type Whittle-Matérn paramétrée par deux réels (c, ν).

L'algorithme converge plutôt rapidement, généralement après 3 ou 4 passages sur les données. Le gain de temps est très important par rapport au ABC classique. En testant plusieurs seuils ε, on voit sur le graphique 1.4 que la loi a posteriori est très corrélée suivant les deux paramètres.

9. le nombre de points observés peut varier d'un bloc à l'autre, ce qui n'est pas le cas ici.



Graphique 1.4 – Ellipses de crédibilité à 50% de la loi a posteriori pour différentes valeurs de ε.

Chapter 2

Introduction (in English)

2.1 Context

Data Science has widely changed over the past few years: the volume of data has been drastically increasing and hardware performance keeps improving. A remarkable example is the Human Genome Project, which was completed in 2003 after years of work. Fifteen years later, it only takes few hours and the cost has been greatly reduced. In 2007, the company Netflix organized a competition with a prize of one million of dollars. The goal was to predict future users' ratings, using a set of preexisting movie ratings. We can see in the Netflix example a reversal in the trend: before, it was common to compute a mean effect. The new data allows us to make individual predictions. In order to accomplish this, we need to develop new tools.

There have been many consequences of this change on statistics as a scientific discipline. Regularization has surfaced as an interesting method to process high dimensional problems. In comparison to other techniques, regularization has shown practical uses with fast algorithms as well as theoretical uses in which performances are close to optimal. This is an active research field as we can see with the recent introduction of a new penalization called the SLOPE in 2015. The classic asymptotic results do not fit the high dimensional setting; we need to develop new non asymptotic tools that require of themselves new proof techniques.

At the same time, improvement of the available computational capacity allows for the development of new inference techniques for models

that were impossible to use before. This is particularly true for Bayesian statistics because they often require numerical methods to compute approximated estimates. The MCMC (Markov Chain Monte Carlo) methods, popularized in the 1990s, are now widely used and process many models. More recently, the ABC (Approximate Bayesian Computation) allows us to tackle a wider class of models, but at the cost of more intensive computer usage.

This thesis fits into this movement: Chapters 3 and 4 study penalized estimators and the particular model of matrix completion. This problem arose quite recently and applies particularly well to individual data. We then try to make a wider analysis: we propose an estimator, study its theoretical properties and then test it on real datasets. Chapter 5 offers an approximated ABC method in the Bayesian setting that is parallelizable. This method is applied to a specific model, showing that it is both quick and efficient.

Since we address it in many chapters, we will start in the next section with some reminders on matrix completion, before describing the three next chapters that are the core of the thesis.

Notations for this chapter

— $\forall n \in \mathbb{N}^*, \quad [n] = \{1, \dots, n\}$.
— $\mathbb{R}^{m_1 \times m_2}$ represents the set of the real $m_1 \times m_2$ matrices.
— For $M \in \mathbb{R}^{m_1 \times m_2}$, $M^\top$ is the transposed matrix.
— $\mathbf{e}_i^m$ is the i-th vector of the $\mathbb{R}^m$ canonic basis, a space that is assimilated to $\mathbb{R}^{m \times 1}$. We write $E_{i,j} = \mathbf{e}_i^{m_1} \mathbf{e}_i^{m_2 \top}$.
— $M_{i,\cdot}$ is the i-th row of M and $M_{\cdot,j}$ is its j-th column.
— $\sigma(M)_k$ represents the k-th singular value of M.
— For $p \geq 1$, we write $\|M\|_{S_p} = (\sum_{k=1}^{\min(m_1,m_2)} \sigma(M)_k^p)^{1/p}$ the p-th Schatten norm of the matrix M. The 2-Schatten norm is also called the *Frobenius norm*.
— The operator norm of M is the largest singular value of M. It is written $\|M\|_{S_\infty}$.
— for any M, we write $\|M\|_\infty$ the maximum of the absolute entries of M.
— $\forall x \in \mathbb{R}$, $\operatorname{sign}(x) = \mathbb{1}\{x > 0\} - \mathbb{1}\{x < 0\}$.
— We write $\mathfrak{m} = \min(m_1, m_2)$ and $\mathfrak{M} = \max(m_1, m_2)$.

2.2 How to complete a matrix?

2.2.1 Global Framework and exact completion

Matrix completion is useful when one wants to guess missing entries that can only be partially observed. To be exact, we will not be presenting the original problem which was addressed by Nathan Srebro with other coauthors (see [Srebro et al., 2005] and [Srebro and Shraibman, 2005] among others).

Without any hypothesis, it is impossible to guess the missing entries. Nevertheless, if the matrix has a particular structure, the problem becomes feasible under certain conditions. The very common hypothesis is that the matrix has a low rank. This hypothesis is quite natural in many real situations: if the rows represent user features, the low rank is obtained when the features are linked and may be summed up by a few components. This hypothesis is related to the PCA [1] (Principal Component Analysis).

Low Rank: a natural hypothesis.

Let us develop an example that may help to visualize the problem and the low rank hypothesis. The data are ratings given to movies by users who watched them. There is a large number of movies as well as of users, but not all of the ratings are known. We can then represent these data as an incomplete matrix where the rows are users and the columns are movies, see Table 2.1.

	The Godfather	Pulp Fiction	Rain Man	OSS117	The Great Dictator	Toy Story
Anna	5	2	.	5	.	.
Pierre	.	5	.	.	.	5
Vincent	1	.	3	5	.	.
Sophie	.	.	.	.	3	3
Keefe	.	1	.	1	.	.
Nicolas	5	.	.	1	5	.

Table 2.1 – Movie ratings as an incomplete matrix (names are fictional)

If it were possible to predict the missing entries, one could thus recom-

1. For PCA, the data are all observed and we only look at a small rank representation.

mend movies to different users based on ratings. We see in this example that individuals' data, which are nowadays easily accessible, allow us to make individualized predictions, whereas in the past we could only focus on the average.

The low rank of the underlying matrix may be analyzed in this example as the similarity among the columns: some users are very similar and only a small number of profile types exist. We can clearly see this by using the Singular Values Decomposition (SVD) of the matrix: any $m_1 \times m_2$ matrix of rank-r can be factorized into a product $U\Sigma V^\top$ where:

— U and V are $m_1 \times r$ and $m_2 \times r$ matrices and the columns are orthonormal;

— Σ is a diagonal matrix and the r positive coefficients are the singular values: $(\sigma(M)_k)_{1\leq k\leq r}$.

Finally, with this deconstruction, we get:

$$M = \sum_{k=1}^{r} \sigma(M)_k U_{\cdot,k} (V_{\cdot,k})^\top.$$

The matrix M is deconstructed into a sum of r rank 1 matrices. Every entry may be written:

$$M_{i,j} = U_{i,\cdot} \Sigma V_{j,\cdot}^\top = \sum_{k=1}^{r} \sigma(M)_k U_{i,k} V_{j,k}.$$

Intuitively, the rating of user i on the j-th movie is the sum of only r components, weighted by the singular values $\sigma(M)_k$. The SVD of M is linked to the eigenvalue deconstruction of $MM^\top$ and $M^\top M$: they both share the same eigenvalues; the eigenvectors of these matrices are respectively the columns of U and V. These operations are the basis of the PCA that can be seen as a SVD of the observations matrix. The presented SVD is a reduced one and it is possible to complete the singular vectors with null singular values.

Conditions for an exact completion.

A central question is the study of the conditions that allow for exact completion. Here, M is a square matrix of size m and rank r; we write $U\Sigma V^\top$ its SVD. The observed locations are denoted by n pairs of integers with value in $[m]\times[m]$: they are gathered in the set Ω. From the SVD, we can see that the rank r matrix has exactly $2mr - r^2$ degrees of freedom:

the matrix Σ has r non null entries, the singular vectors on the left have $(m-1)+\cdots+(m-r) = mr - r(r+1)/2$ degrees of freedom and so on for the right singular vectors. If $n < 2mr - r^2$, many matrices may correspond to the observed entries and the solution is not unique among the rank r matrices.

We can see that $2mr - r^2$ is a lower bound of the number of observations. A natural question thus arises: what is the minimal sample size needed to identify exactly the matrix when entries are randomly selected, and under which assumptions? In a first attempt, it is easier to deal with a uniform sampling: every sample of size n have the same probability.

Another consideration is that some particular matrices are more difficult to reconstruct than others. For example, a matrix with 0 everywhere except in one position is almost impossible to be reconstructed except if the non null entry is in the sample. For the following matrix:

$$M = \mathbf{e_1^m e_m^{m\top}} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{2.1}$$

the upper right entry needs to be observed if one wants to reconstruct the matrix, which happens with probability n/m^2. Otherwise it is impossible to guess the place and the value even if we know that M is rank 1. The required assumptions concern the singular vectors of M: their entries have to be spread out and not too uneven. In our last example, the singular vectors of the matrix (2.1) are from the canonical basis so it is the complex case.

The rank relaxation.

This is not actually military jargon, but instead involves the reconstruction program. Intuitively, there are many matrices that correspond to the observed values. In order to find one with a low rank, the seminal program is thus:

$$\begin{array}{rl} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & \forall (i,j) \in \Omega,\, X_{i,j} = M_{i,j} \end{array} \tag{2.2}$$

The function to be optimized is not convex and is therefore problematic. A fundamental idea at this stage is to replace the rank function by a convex relaxation, which is for instance the nuclear norm. This norm,

also known as the Schatten-1 norm, is a convex surrogate of the rank. It is expressed for a matrix M as:

$$\|M\|_{S_1} = \sum_{k=1}^{r} \sigma(M)_k.$$

Due to the convexity of a norm, the following program:

$$\begin{array}{ll} \text{minimize} & \|X\|_{S_1} \\ \text{subject to} & \forall (i,j) \in \Omega,\, X_{i,j} = M_{i,j} \end{array} \tag{2.3}$$

is much easier to solve. It may be seen as a program of the SDP class (*Semidefinite Programing*). To our knowledge, this idea was proposed for the first time by Fazel et al. [2001]. The nuclear norm was subsequently used widely for many close problems. We now have at hand every ingredients we need to state the main theorems for exact matrix completion.

Theorem 2.1 (Theorem 1.2 from [Candès and Tao, 2010])**.** *Let $M \in \mathbb{R}^{m_1 \times m_2}$ be a fixed matrix of rank r obeying the* strong incoherence property [2] *with parameter μ. Suppose we observe n entries of M with locations sampled uniformly at random. Then there exists a positive numerical constant C such that if:*

$$n \geq C\mu^2\, \mathfrak{M}\, r (\log \mathfrak{M})^6,$$

M is the unique solution to (2.3) *with probability at least $1 - n^{-3}$. In other words: with high probability, nuclear norm minimization recovers all the entries of M with no error.*

This result improves a former article [Candès and Recht, 2012] where the minimal amount of observed entries is of order $\mathfrak{M}^{1.2}\, r \log \mathfrak{M}$ to allow a reconstruction with high probability. It is an important and surprising result: up to a logarithm term, one only needs of order $\mathfrak{M}\, r$ entries in order to reconstruct the whole matrix, that may be much lower than $m_1 m_2$ if the rank is small. It almost reach the minimum number of entries needed in the best case for a uniform random draw.

2.2.2 Completion with noise by least squares

The previously proposed problem is quite particular because the randomness is only involved in the sampling. The question of an exact

2. The precise definition is in the article. The goal is to avoid pathological cases as the one presented in (2.1).

reconstruction may be seen as a little bit excessive in the sense that, in many situations, it is sufficient to obtain a rather good idea and a standard error. Furthermore, the noiseless observation framework is not very realistic. By the way, the noisy case is more suitable to the statistician because there is a parameter to estimate. Note that in the noisy case, the matrix corresponding to the observations is no longer of low rank so the tools have to be adapted. In the following, we consider $m_1 \times m_2$ matrices.

The *Trace* regression.

We introduce here the *Trace* regression model, which encompasses the matrix completion. Let $\langle \cdot \rangle$ be the canonic scalar product over the $m_1 \times m_2$ real matrices:

$$\langle A, B \rangle = \mathrm{Tr}(A^\top B) = \mathrm{Tr}(B^\top A).$$

The associated norm is the Frobenius norm, which is also the Schatten-2 norm:

$$\|A\|_{S_2} = \sqrt{\mathrm{Tr}(A^\top A)} = \sqrt{\sum_{i,j} A_{i,j}^2}.$$

The statistical framework is as follows: for $i \in \{1, \dots, n\}$, we observe $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^{m_1 \times m_2}$ from the model:

$$Y_i = \langle M^\star, X_i \rangle + \sigma \varepsilon_i,$$

where $M^\star$ is the parameter to be estimated and (ε_i) is an independent and identically distributed sequence of standard noise.

The matrix completion task is a particular case when the design matrices (X_i) are *mask* matrices: there is a 1 at one location and it is null otherwise. Hence they are elements of the set:

$$\mathcal{X} = \left\{ {\mathbf{e}_j^{m_1}}^\top \mathbf{e}_k^{m_2},\ (j,k) \in [m_1] \times [m_2] \right\}.$$

For a mask matrix where the entry at (k,l) is not null (if $X_i = E_{k,l}$), we get that for any matrix M: $\langle M, X_i \rangle = M_{k,l}$.

The common estimator is a penalized least squares estimator. For $\lambda \in \mathbb{R}$, it is defined by:

$$\widehat{M} = \operatorname*{arg\,min}_{M \in \mathcal{C}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle M, X_i \rangle)^2 + \lambda \|M\|_{S_1} \right\}, \tag{2.4}$$

where $\mathcal{C}$ is a subset, usually convex, of $\mathbb{R}^{m_1 \times m_2}$.

Many works have studied the theoretical properties of this estimator with different conditions on noise, design and underlying matrix $M^\star$. After the works [Bach, 2008], [Candès and Plan, 2010], [Keshavan et al., 2010], [Rohde and Tsybakov, 2011] and [Negahban and Wainwright, 2012] (among others), we cite here a result from [Klopp, 2014], which deals with a rather general framework.

Main result from [Klopp, 2014].

Two conditions involve the sampling of the locations. Every entry and also the rows and the columns have to be sampled with a not too small probability: the uniform sampling is not mandatory but we can not process extreme sampling. The matrices X_i are i.i.d. drawn from a distribution Π in $\mathcal{X}$. If we write $\pi_{j,k}$ the probability of drawing $E_{j,k}$, the conditions are as follows:

Assumption 2.1. *There exists a constant $L > 1$ such that:*

$$\left.\begin{array}{c} \max_j \left(\sum_{k=1}^{m_2} \pi_{j,k}\right) \\ \max_k \left(\sum_{j=1}^{m_1} \pi_{j,k}\right) \end{array}\right\} \leq L/\mathfrak{m}\,.$$

Assumption 2.2. *There exists a constant $\mu \geq 1$ such that, for every location (j,k):*

$$\pi_{j,k} \geq (\mu m_1 m_2)^{-1}.$$

For a uniform sampling, we get $L = \mu = 1$. The last assumption involves the noise which has to be subexponential.

Assumption 2.3. *There exists $K > 0$ such that:*

$$\max_{i \in \{1,\dots,n\}} \mathbb{E} \exp\left(|\varepsilon_i| / K\right) < \infty.$$

It is now possible to state the main theorem.

Theorem 2.2 (Theorem 7 from [Klopp, 2014])**.** *Let (X_i) be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumptions 2.1 and 2.2. Assume that*

$\|M^\star\|_\infty \leq \boldsymbol{a}$ *and that Assumption 2.3 holds. Consider the regularization parameter* λ *satisfying:*

$$\lambda = C\sigma\sqrt{\frac{L\log(m_1+m_2)}{n\,\mathfrak{m}}},$$

where C *is a known constant. Let* $\widehat{M}$ *be defined in* (2.4) *where* $\mathcal{C} = \{M : \|M\|_\infty \leq \boldsymbol{a}\}$*. Then, there exists a numerical constant* c'*, that depends only on* $(K, \sigma, \boldsymbol{a}, \mu, L)$ *such that:*

$$\frac{\left\|\widehat{M} - M^\star\right\|_{S_2}^2}{m_1 m_2} \leq c' \max\left\{\frac{\operatorname{rank}(M^\star)\,\mathfrak{M}\log(m_1+m_2)}{n}, \sqrt{\frac{\log(m_1+m_2)}{n}}\right\}.$$

The parallel to the LASSO estimator is clear. The nuclear norm replaces the ℓ_1 norm, and is actually the ℓ_1 norm applied to the vector of the singular values. The size of the parameter is of order $r\,\mathfrak{M}$ and the upper bound is of order $r\,\mathfrak{M}/n$ up to log terms. We will see in Chapter 4 that a unified framework exists, following ideas from [Lecué and Mendelson, 2015a,b].

Different results of lower bounds state that the minimax rate is of order $\operatorname{rank}(M^\star)\,\mathfrak{M}/n$. The given upper bound is therefore optimal up to a logarithm term. The dependence on the noise may be avoided by considering a different estimator, which is very similar to the square-root Lasso, by taking the square root of the adjustment term. This estimator is also considered in the same article. In practice, the cross-validation is used in order to tune λ. In a slightly different model, the same author reaches to delete the logarithm term, see [Klopp, 2015], and get the minimax rate.

The article [Koltchinskii et al., 2011] treats a more general framework and not only the mask matrices for the design. The proposed estimator is slightly different and the design has to be known even if it does not have to be uniform.

2.2.3 Some Algorithms

It is important to us to explain some points about the computation of the estimator even though it is not the core of the thesis. The function to minimize in (2.4) is convex but some issues remain. Here, we must first

recall two properties about the nuclear norm. The sub-differential of the nuclear norm (see [Watson, 1992]) at M whose SVD is written $U\Sigma V^\top$ is:

$$\partial \left\|\cdot\right\|_{S_1}(M) = \left\{UV^\top + W, \quad \left\|W\right\|_{S_\infty} \leq 1,\, U^\top W = 0, WV = 0\right\}.$$

When the matrix M is low rank, the sub-differential of the nuclear norm is wide because there are many possibilities for W. A function that is very important is the soft-thresholding operator. We define it, for $\lambda > 0$ by:

$$S_\lambda(M) = UDV^\top \quad \text{where} \quad D = \operatorname{diag}\left(\{(\max(0, \sigma_k(M) - \lambda)\}_{1 \leq k \leq r}\right).$$

Similarly to the vectors, we get a result involving the proximal operator of the nuclear norm and the soft-thresholding of the singular values S_λ.

Proposition 2.1. *Let M be a $m_1 \times m_2$ matrix. Therefore:*

$$\arg\min_X \left\{\frac{1}{2}\left\|X - M\right\|_{S_2}^2 + \lambda \left\|X\right\|_{S_1}\right\} \ni S_\lambda(M).$$

Proof, following [Recht et al., 2010]. We set the function h defined for any matrix X by $h(X) = \frac{1}{2}\left\|X - M\right\|_{S_2}^2 + \lambda \left\|X\right\|_{S_1}$. The function h is strictly convex as a sum of two strictly convex functions so the minimum is unique. X_0 minimizes h if:

$$0 \in \partial h(X_0) = \left\{X_0 - M + \lambda D \,:\, D \in \partial \left\|\cdot\right\|_{S_1}(X_0)\right\}.$$

We will show that the last equation holds for $X_0 = S_\lambda(M)$. We write U_0, V_0 the singular vectors associated to singular values of M that are greater than λ; U_1, V_1 for the singular values lower than λ; Σ_0 and Σ_1 are the two diagonal corresponding matrices. Hence we have:

$$X_0 = S_\lambda(M) = U_0(\Sigma_0 - \lambda I)V_0^\top.$$

In order to make it work, we set:

$$W_0 = \lambda^{-1} U_1 \Sigma_1 V_1^\top.$$

We see that $\left\|\lambda^{-1}U_1\Sigma_1V_1^\top\right\|_{S_\infty} \leq 1$ because the entries of Σ_1 are bound by λ. Furthermore, $U_0^\top W_0 = W_0 V_0 = 0$ because the vectors of U_1 are

orthogonal to the U_0 ones and similarly for V_1 and V_0. Finally, $U_0V_0^\top + W_0$ is in $\partial \|\cdot\|_{S_1}(M_0)$.

After picking up a relevant element of the sub-differential at X_0, we can conclude because:

$$X_0 - M + \lambda\left(U_0V_0^\top + W_0\right) = 0.$$

■

Proposition 2.1 actually deals with a problem close to the matrix completion issue. This may be seen as a denoising task where each entry is observed with noise and we seek a low rank representation [3]. This proximal operator is at the basis of many algorithms that perform matrix completion, the first one of a long list being [Cai et al., 2010]. The proximal operator S_λ is also used for ADMM-type algorithms (Alternating Direction Method of Multipliers), which are quite popular for matrix completion.

We propose Algorithm 2.1, that follows the scaled form from [Boyd et al., 2011]. The goal is to compute the minimizer without the constraint:

$$\widehat{M} = \underset{M\in\mathbb{R}^{m_1\times m_2}}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle M, X_i\rangle)^2 + \lambda\|M\|_{S_1}\right\},$$

by splitting the problem into two parts that are identified in the algorithm by M and N.

Algorithm 2.1 ADMM Algorithm for matrix completion

Initialization $\varepsilon, M^0, N^0, U^0, t = 0$.
While $\left\|U^t - U^{t-1}\right\|_{S_2} > \varepsilon$, do:

1. $t \leftarrow t+1$
2. $M^t \leftarrow \arg\min_M \left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle M, X_i\rangle)^2 + \frac{\rho}{2}\left\|M - N^{t-1} + U^{t-1}\right\|_{S_2}\right\}$
3. $N^t \leftarrow \arg\min_N \left\{\lambda\|N\|_{S_1} + \frac{\rho}{2}\left\|M^t - N + U^{t-1}\right\|_{S_2}\right\}$
4. $U^t \leftarrow U^{t-1} + M^t - N^t$

Return M^t.

In this algorithm, the update of M is entrywise so it is cheap. On the other hand, the update of N needs a SVD by using the proximal

3. the PCA would be the equivalent to the hard-thresholding operator.

operator. The bottleneck of these algorithms is the computation of the SVD, which is very costly (of order $\mathfrak{m}^2\,\mathfrak{M}$). Other methods therefore appeared in order to bypass this issue.

The first possibility is to compute an approximated SVD: the values and vectors are not exactly computed and a small tolerance is allowed; there exist stochastic methods for this. In the same spirit, some algorithms only compute the N largest singular values and the associated singular vectors, for a fixed N. As S_λ only requires the singular values greater than λ, one can set a small value for N and increase it if the last singular value is above the threshold. This method is quite difficult to tune in practice but the routines to compute the N largest singular values are implemented in FORTRAN and are therefore pretty fast.

An associated bi-convex problem.

The approximated SVD methods are not efficient enough for a very large scale. A more important problem is that it does not really take into account the low-rankness and the ability of a matrix to be summed up by few components: it needs to store a $m_1 \times m_2$ matrix. Here, one can use a factorization that comes from the expression of the nuclear norm:

$$\forall M \in \mathbb{R}^{m_1 \times m_2}, \quad \|M\|_{S_1} = \frac{1}{2} \min_{\substack{LR^\top = M \\ L \in \mathbb{R}^{m_1 \times \mathfrak{m}} \\ R \in \mathbb{R}^{m_1 \times \mathfrak{m}}}} \left\{ \|L\|_{S_2}^2 + \|R\|_{S_2}^2 \right\}.$$

By using the SVD of $M = U\Sigma V^\top$ and taking $L_0 = U\Sigma^{1/2}, R_0 = V\Sigma^{1/2}$, we immediately get $\|X\|_{S_1} = \|L_0\|_{S_2}^2 + \|R_0\|_{S_2}^2$. To show the other side of the inequality, the SDP characterization of the nuclear norm can be used. By a double minimization, we get that the estimator (2.4) is therefore equal to $\widehat{L}\widehat{R}^\top$ where:

$$(\widehat{L}, \widehat{R}) = \operatorname*{arg\,min}_{\substack{L \in \mathbb{R}^{m_1 \times \mathfrak{m}} \\ R \in \mathbb{R}^{m_1 \times \mathfrak{m}}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle LR^\top, X_i \rangle)^2 + \frac{\lambda}{2} \left[\|L\|_{S_2}^2 + \|R\|_{S_2}^2 \right] \right\}. \tag{2.5}$$

This program is more appealing because it only involves quadratic terms that are differentiable. The problem is convex in both L and R separately, so it is called bi-convex. Unfortunately, it is not convex with respect to the pair (L, R): it means that there are many local minima. Nevertheless, it has been proposed by [Mazumder et al., 2010] and a

distributed version may be found in [Recht and Ré, 2013]. The theoretical work [Burer and Monteiro, 2003] shows that the problem (2.5) is not so bad and the solution would be unique [4].

Furthermore, by taking an upper bound of the rank of the underlying matrix, it is possible to decrease the storage requirement. This remark will be developed in the Bayesian section because it uses a similar factorization.

2.2.4 Binary matrix completion

After the problem of least squares matrix completion, statisticians turned to other close problems. For instance, we may consider prediction when the entries are binary. This happens often on Internet websites where users can choose between two options, such as I like it and I don't like it. This problem, which we may call *classification* as opposed to *regression*, needs new tools. We label the observed entries in $\{-1, +1\}$.

A common way to model a classification problem is to assume a generalized regression model. The interpretation of this model is made easier by incorporating a latent variable for each observation $Y_i^\star$. It can be seen, such as in the case of movie ratings, as a continuous measure of the taste of that individual for a film. The observation is then $+1$ if it is above 0 and -1 otherwise. Mathematically, it is written as:

$$Y_i^\star = \langle M^\star, X_i \rangle + Z_i,$$

where Z_i has a symmetric distribution F. We then observe:

$$Y_i = \mathbb{1}\{Y_i^\star \geq 0\} - \mathbb{1}\{Y_i^\star < 0\}.$$

The direct distribution of Y_i is:

$$Y_i = \begin{cases} +1 & \text{with probability } \mathbb{P}(Z_i \geq -\langle M^\star, X_i \rangle) = F(\langle M^\star, X_i \rangle) \\ -1 & \text{with probability } \mathbb{P}(Z_i < -\langle M^\star, X_i \rangle) = F(-\langle M^\star, X_i \rangle) \end{cases} \tag{2.6}$$

The normalized log-likelihood, which will be used as the adjustment term, is:

$$L(M) = \frac{1}{n} \sum_{i=1}^{n} \log F(Y_i \langle M, X_i \rangle).$$

The usual choices for F are:

4. The uniqueness is not about (L, R) but about the product.

— the logistic function, $F(x) = (1 + \exp(-x))^{-1}$
— the Gaussian cumulative distribution function, usually written Φ.

This model has been studied by many authors. The first work [Davenport et al., 2014] states the binary regression model. The sampling is uniform [5] and the proposed estimator is the solution of the program:

$$\begin{array}{ll} \text{Maximize} & L(M) \\ \text{subject to} & \|M\|_{S_1} \leq \alpha\sqrt{sm_1m_2} \text{ and } \|M\|_\infty \leq \alpha. \end{array} \tag{2.7}$$

The constant s is not the rank but a parameter related to the nuclear norm. We see that the knowledge of s is needed in order to compute the estimate. The theoretical results in this work do not involve low rank matrices but low nuclear norm matrices. There are some technical assumptions over F that involves the smoothness of the function. The main result is the reconstruction of the matrix $M^\star$ and is the following.

Theorem 2.3 (Theorem 1 from [Davenport et al., 2014])**.** *Assume that* $\|M^\star\|_{S_1} \leq \alpha\sqrt{sm_1m_2}$ *and* $\|M^\star\|_\infty \leq \alpha$*. Suppose that the observations are generated as in* (2.6) *with a uniform sampling. Consider the estimator* $\widehat{M}$ *from* (2.7)*. Then, if* $n \geq (m_1 + m_2)\log(m_1m_2)$*, with high probability:*

$$\frac{1}{m_1m_2}\left\|\widehat{M} - M^\star\right\|_{S_2}^2 \leq C\sqrt{\frac{s(m_1 + m_2)}{n}},$$

where C *is an absolute constant that only depends on* F *and* α*.*

At the first glance the rate is with a square root that does not look very good. It comes from the considered class of matrices (the matrices that have a bounded nuclear norm), which is wider than the low rank matrices. In this class, the rate is almost optimal and we will see this rate in Chapter 4. The most important problem in this result is that we do not recover the sparsity induced by the nuclear norm regularization.

This work is done in a sequence of articles [Lafond et al., 2014, Klopp et al., 2015]. There is a link made between the nuclear norm and the rank, and the rate for reconstructing the matrix is similar to the one in the regression problem by least squares.

We should recall here the main theorem without stating the technical conditions about the smoothness of F. The estimator is the one where

5. Formally, it is a different sampling model, which is similar to the one in [Klopp, 2015].

the log-likelihood substitutes the quadratic adjustment term:

$$\widehat{M} = \underset{\|M\|_\infty < \mathbf{a}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log F(Y_i \langle M, X_i \rangle) + \lambda \|M\|_{S_1} \right\} \tag{2.8}$$

Theorem 2.4 (Corollary 2 from Klopp et al. [2015]). *Let $(X_i, Y_i)_{1 \le i \le n}$ be an i.i.d. sample from (2.6) where (X_i) is drawn from the distribution Π on $\mathcal{X}$ and satisfies Assumptions 2.1 and 2.2. Let $M^\star$ be a real matrix such that $\|M^\star\|_\infty \le \boldsymbol{a}$. The regularization parameter λ is set:*

$$\lambda = C\sigma \sqrt{\frac{\log(m_1 + m_2)}{\mathfrak{m}\, n}},$$

where C is a constant that depends on L and F. We consider the estimator $\widehat{M}$ satisfying (2.8). Then, there exists an absolute constant c' that depends only on $(F, \boldsymbol{a}, \mu, L)$ such that:

$$\frac{\left\|\widehat{M} - M^\star\right\|_{S_2}^2}{m_1 m_2} \le c' \max \left\{ \frac{\operatorname{rank}(M^\star)\, \mathfrak{M} \log(m_1 + m_2)}{n}, \sqrt{\frac{\log(m_1 + m_2)}{n}} \right\}.$$

with probability greater than $1 - 3/(m_1 + m_2)$.

We recognize here the usual rate of order $\operatorname{rank}(M^\star)\, \mathfrak{M} \log(m_1 + m_2)/n$. The authors exhibit a lower bound of order $\operatorname{rank}(M^\star)\, \mathfrak{M}\, /n$, that meets the upper bound up to a logarithmic factor.

Another point of view of the 1-bit matrix completion.

If this result on the reconstruction of $M^\star$ is optimal, it assumes a generalized regression model. Moreover, is the reconstruction of the matrix the most interesting criterion? In a classification task, we may suppose that it is more important to control the misclassification rate in prediction. Formally, for a draw X_i, the 0/1 prediction risk for any matrix M is therefore:

$$R_{0/1}(M) = \mathbb{P}[\operatorname{sign}(\langle M, X_i \rangle) \neq Y_i] = \mathbb{E}[\mathbb{1}\{\operatorname{sign}(\langle M, X_i \rangle) \neq Y_i\}]. \tag{2.9}$$

An interesting and complete study of many different loss functions can be found in [Zhang, 2004]: this article makes the link between the risk induced by loss functions used in practice, which are often the convex

surrogate of the 0/1 loss, and the 0/1 risk. The article shows that the upper bound for the logistic risk is transferred with a square root to the 0/1 risk. It is therefore not optimal to use it and leads to the study of other loss functions. We can expect to get an optimal 0/1 risk of classification; it is a large motivation for Chapters 3 and 4.

2.3 Bayesian framework - completion by squared loss

We can call the previous approach *frequentist*. The *Bayesian* approach is also suitable to the matrix completion issue but has to be adapted. The framework that is presented here, because it will be used in Chapter 3, is rather called PAC-Bayesian but the differences are small. Roughly, by denoting θ the parameter, the PAC-Bayesian estimation uses an empirical loss function written $r(\theta)$ that incorporates the data (the dependency to the data is implicit) and a prior distribution written $\pi(\theta)$. We therefore define the pseudo posterior distribution:

$$\widehat{\rho}_\tau(d\theta) = \frac{\exp[-\tau r(\theta)]}{\int \exp[-\tau r] d\pi} \pi(d\theta). \tag{2.10}$$

The positive number τ is called the inverse temperature and plays the opposite role of λ in the penalized estimator (2.4). The PAC-Bayesian estimator is then computed as the expectation with respect to the pseudo posterior distribution:

$$\widehat{\theta}_\tau = \int \theta \widehat{\rho}_\tau(d\theta).$$

Parallel to the previous estimators, the prior distribution is analog to the regularization by favoring the places where it is charged. The estimator is a mean rather than an isolated one. The loss function that is usually used for the matrix completion is the square loss function: $r(M) = 1/n \sum_{i=1}^{n} (Y_i - \langle X_i, M \rangle)^2$. It corresponds to a Gaussian noise.

A difficulty that arises at this stage is to build a prior distribution that favors low rank matrices. In order to do that and following the first Bayesian articles [Lim and Teh, 2007, Salakhutdinov and Mnih, 2008], we use the factorization into a product of two matrices. A matrix M of rank $K \leq \mathfrak{m}$ can be indeed written by a product of two matrices $L \in \mathbb{R}^{m_1 \times K}$ and $R \in \mathbb{R}^{m_2 \times K}$ such as:

$$M = LR^\top.$$

This factorization is obviously not unique. If the rank r of M is strictly lower than K, the factors may have exactly r non null columns. On the opposite the pairs (L, R), with exactly K columns, may generate any matrix with a rank lower than K. The choice of K has to be done by the practitioner but, for theoretical results, we can take the most general value that is $\mathfrak{m}$. The important fact is that the results do not depend on the choice of K: they have to be rank adaptive and it is the case of the next results. A large value involves the storage of larger parameters but allows to reconstruct a matrix with a larger rank. In the following we take $K = \mathfrak{m}$ even though in practice it is common to take a lower value.

The low rank is, in this framework, induced by a prior distribution that favors null columns. These ideas are quite similar to the Group Lasso introduced by Yuan and Lin [2006]; a Bayesian model is studied in [Kyung et al., 2010]. The basic idea is to use a hierarchic prior with an additional parameter $\gamma = (\gamma_k)_{1 \leq k \leq \mathfrak{m}}$. The prior distribution over L, R is a distribution centered around 0. The dispersion, indexed by γ_k, is specific to the column; the aim of this construction is to shrink a whole column to 0. The hierarchical model is flexible by allowing some columns to have a large dispersion and some others a very low dispersion and therefore almost collapse. Formally, the prior distribution is:

$$\begin{aligned} \forall (i,k) \in [m_1] \times [\mathfrak{m}], \quad & L_{i,k} | \gamma_k \underset{i.i.d}{\sim} \mathcal{N}(0, \gamma_k) \\ \forall (j,k) \in [m_2] \times [\mathfrak{m}], \quad & R_{j,k} | \gamma_k \underset{i.i.d}{\sim} \mathcal{N}(0, \gamma_k) \\ \forall k \in [\mathfrak{m}], \quad & \gamma_k \underset{i.i.d}{\sim} \pi^\gamma \end{aligned} \tag{2.11}$$

where π^γ is a prior distribution on $\mathbb{R}^+$ that is clarified when needed. The conjugate model is obtained by using a Inverse-Gamma distribution as usual for the variance of a Gaussian distribution.

The parameter is eventually: $\theta = (L, R, \gamma) \in (\mathbb{R}^{m_1 \times \mathfrak{m}}) \times (\mathbb{R}^{m_2 \times \mathfrak{m}}) \times \mathbb{R}^{\mathfrak{m}}$. The estimated matrix is therefore:

$$\widehat{M}_\tau = \int L R^\top \widehat{\rho}_\tau(d\theta). \tag{2.12}$$

We may study the properties of this matrix by comparing its risk to the one from the oracle $M^\star$. It is the *machine learning* approach in the sense that we do not assume a statistical model. We also study the problem of computing $\widehat{\theta}_\tau$: the posterior distribution is not explicit, as it often is in a quite complex model. We will see two methods to approximate the estimate.

2.3.1 Theoretical guarantees of the Bayesian estimator.

The Tien Mai and Pierre Alquier have studied the properties of the PAC-Bayesian estimator in [Mai and Alquier, 2015].We present here their main result. We assume that the observations are generated according to the following model:

$$Y_i = \langle M^\star, X_i \rangle + \varepsilon_i, \tag{2.13}$$

where $(\varepsilon_i)_{1\leq i\leq n}$ is an i.i.d. sequence of centered noise and $(X_i)_{1\leq i\leq n}$ are i.i.d. taking value in $\mathcal{X}$ following the distribution Π. The knowledge of the noise distribution is not required here. The goal is to control the gap between $M^\star$ and $\widehat{M}_\tau$. The authors look at the weighted Frobenius norm:

$$R(M) = \|M - M^\star\|^2_{S_2,\Pi} = \sum_{j,k} \pi_{j,k}(M - M^\star)^2_{j,k} = \mathbb{E}[\langle X_i, M - M^\star\rangle^2].$$

We do not need any assumption on the sampling distribution, denoted by Π. If Assumption 2.2 is verified for any $\mu \geq 1$, so if every entry has a positive probability to be drawn, we get the following inequality that links the two norms:

$$\|M - M^\star\|^2_{S_2,\Pi} \geq \frac{1}{\mu}\frac{1}{m_1 m_2}\|M - M^\star\|^2_{S_2}.$$

For technical reasons, the prior distribution over the entries of L and R has to be bounded: the authors then use a uniform distribution that depends on the activation of the column (we set $\delta \gg \kappa$):

$$L_{i,k}, R_{j,k}|\gamma_k \sim \begin{cases} \mathcal{U}_{[-\delta,\delta]} & \text{si } \gamma_k = 1 \\ \mathcal{U}_{[-\kappa,\kappa]} & \text{si } \gamma_k = 0 \end{cases}$$

The prior distribution on the vector γ is also simple:

$$\gamma = (\underbrace{1,\ldots,1}_{k \text{ times}}, \underbrace{0,\ldots,0}_{\mathfrak{m}-k \text{ times}}) \text{ with probability } \frac{\beta^{k-1}(1-\beta)}{1-\beta^{\mathfrak{m}}}.$$

We finally need an assumption about the noise, that has to be subexponential as in Theorem 2.2.

We can now state the main theorem by introducing the set $\mathcal{M}(k)$, a subset of matrices whose rank is at most k (α is a positive number):

$$\mathcal{M}(k) = \left\{LR^\top, (L,R) \in \mathbb{R}^{m_1\times k} \times \mathbb{R}^{m_2\times k}, (\|L\|_\infty, \|R\|_\infty) \leq \alpha/\mathfrak{m}\right\}.$$

Theorem 2.5 (Theorem 1 from [Mai and Alquier, 2015])**.** *Assume that the observations $(X_i, Y_i)_{1\leq i\leq n}$ are generated from (2.13) and that Assumption 2.3 holds. Assume that $\|M^\star\|_\infty \leq \boldsymbol{a}$ and take $\tau = n/2C$, δ, κ being fixed. The estimator $\widehat{M}_\tau$ is defined in (2.12). Then, for any $\epsilon \in]0,1[$, as soon as $n > \mathfrak{M}$, with probability greater than $1-\epsilon$ we get:*

$$\left\|\widehat{M}_\tau - M^\star\right\|_{S_2,\Pi} \leq \min_{k\in[\mathfrak{m}]} \left\{ 3 \inf_{M\in\mathcal{M}(k)} \|M - M^\star\|_{S_2,\Pi} + C_2 \frac{(m_1+m_2)k\log\mathfrak{m}}{n} + \frac{C_3 \log\frac{2}{\epsilon}}{n} \right\},$$

where C, C_2, C_3 are constants that only depend on L, β and the noise.

As opposed to the frequentist results, it is an oracle inequality in the sens that it does not require $M^\star$ to be low rank. If this matrix is full rank but can be well approximated by a low rank matrix (for instance when its singular values decrease quickly), the rate of convergence is the best trade-off between bias and variance. This rate is optimal up to a logarithm term. Moreover, the assumptions about the sampling are minimal and it is easy to make the assumptions in order to bound the Frobenius norm. The issue of the computation of the estimator is still open. The first method explained later, called MCMC, has been implemented in the aforementioned article and it works well on simulated datasets that are not too large, say a 100×100 matrix but it is prohibitive for very large matrices. We will see after that an approximated method that is much faster.

2.3.2 Computation of the Bayesian estimator: MCMC

The most popular way to compute an approximation of the estimator is to use MCMC, *Markov Chain Monte Carlo*. The idea is to draw a large sample from a Markov chain whose invariant distribution is the posterior distribution. If we write this sample of size T: $(\theta^t)_{1\leq t\leq T}$, the estimator is approximated by:

$$\tilde{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \theta^t.$$

If the chain has the required properties, the ergodic theorem states that $\tilde{\theta}_T$ goes to $\widehat{\theta}_\tau$ when T goes to $+\infty$. The interested reader is referred to [Robert and Casella, 2005] for a long and complete introduction to the MCMC methods.

Among several methods, the matrix completion model as presented here is quite easy to deal with because the likelihood corresponds to a Gaussian noise and the prior distribution of L and R as well. It is easy to derive the conditional distributions in closed form (by taking π^γ in a smart way) and a Gibbs sampler is then built. It is done, with slightly modifications, in [Salakhutdinov and Mnih, 2008]. The model is applied on Netflix prize data and performs well. Nevertheless, with regard to the size of the data and the dimension of the matrix, this method is very time consuming and is at its limit. Moreover, the theoretical guarantees are very complex in order to prove the convergence of the chain.

2.3.3 Computation of the estimator: variational approximation

It is very common in mathematics, for a function that is hard to compute, to seek an approximation. This idea is on the basis of the so-called variational Bayes approximation (denoted by VB) where the posterior distribution $\widehat{\rho}_\tau$ is fully approximated by a function whose moments are easier to compute (for a good introduction, the reader may read [Bishop, 2006], chapter 10; the method is very popular and has been presented in a NIPS tutorial in 2016). In order to define the approximation, we need a dissimilarity measure and a function family. For the dissimilarity, we use here the Kullback-Leibler divergence, even though other functions suit. If μ is a measure that is absolutely continuous with respect to ν, we define the Kullback-Leibler divergence $\mathcal{K}(\mu, \nu)$ by:

$$\mathcal{K}(\mu, \nu) = \int \log \frac{d\mu}{d\nu} d\mu.$$

Given a function family $\mathcal{F}$, we then seek:

$$\widetilde{\rho}_\tau = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \widehat{\rho}_\tau). \tag{2.14}$$

The approximated estimator, that depends on the family $\mathcal{F}$, is therefore the mean with respect to this distribution:

$$\widetilde{\theta}_\tau = \int \theta \widetilde{\rho}_\tau(d\theta).$$

The family $\mathcal{F}$ has to be small for the ease of the computation of the minimizer (and also the expectation) but the family has to be as large

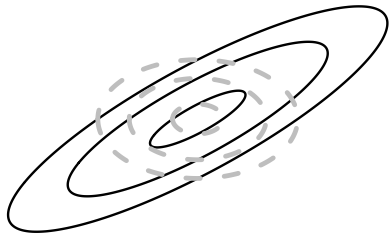


Figure 2.1 – Approximation of a bivariate normal distribution with 0.85 correlation by independent components.

as possible in order to give a good approximation. In the following, we will only deal with the case where all the distributions are absolutely continuous with respect to a reference measure.

A first popular class of families in the literature is the family of independent components (this class is usually called *Mean Field*). Given a writing of the parameter in N blocks: $(\theta_i)_{i\in[N]}$, the family is:

$$\mathcal{F}^{MF} = \left\{ q : q(\theta) = \prod_{i=1}^{N} q_i(\theta_i) \right\}.$$

Intuitively, the joint distribution $\widehat{\rho}_\tau$ is approximated by a distribution that has independent components. Graph 2.1 shows the approximation of a bivariate normal distribution (with correlated components) by two independent Gaussian distributions. From the factorization, it comes that $\widetilde{q}$ is a fixed point satisfying:

$$\forall i \in [N],\, q_i(\theta_i) \propto \exp\left(\int \{-\tau r(\theta) + \log \pi(\theta)\} \prod_{j\neq i} q_j(\theta_j) d\theta_j \right).$$

If the prior distribution is appropriately chosen, the densities (q_j) are parametric and the optimization is equivalent to seek a fixed point in a finite dimension parameter.

Another possibility is a parametric family. We write:

$$\mathcal{F}^{P} = \{ f_m : m \in \mathcal{M} \},$$

where $\mathcal{M}$ has a finite dimension. In order to compute the optimal element

in practice, we use the following identity:

$$\widetilde{\rho}_\tau = \underset{\rho\in\mathcal{F}}{\arg\min}\, \mathcal{K}(\rho, \widehat{\rho}_\tau) = \underset{\rho\in\mathcal{F}}{\arg\min} \left\{ \int \tau r(\theta)\rho(d\theta) + \mathcal{K}(\rho, \pi) \right\}. \qquad (2.15)$$

The right member of (2.15) may be easy to compute, or may be bounded by a quantity that can be easily optimized. The reader is referred to [Alquier et al., 2016] for wider explanations. This article also gives a way to compute theoretical bounds of the risk of the estimator. Similar ideas will be used in Chapter 3.

The matrix completion model with the quadratic loss has been treated with this method with a *Mean Field* approach by Lim and Teh [2007]. As all the distributions are well chosen, the calculus is direct and explicit and the algorithm is therefore built. It converges quickly on a large dataset. The problem at this stage is that we do not have any guarantees about the convergence of the algorithm nor about the quality of the approximation: it is not the true Bayesian estimator. In Chapter 3, we will develop a variational estimator for binary matrix completion and we are able to derive theoretical properties about the quality of the approximation.

2.4 Summary of the Chapters

2.4.1 Chapter 3: *1-bit Matrix Completion: PAC-Bayesian Analysis of a Variational Approximation*

This chapter proposes an estimator for the binary matrix completion issue, where the observed entries lie in $\{-1, +1\}$. The former works, that use a generalized linear model, are not completely conclusive because they focus on the reconstruction of the matrix parameter and not on the prediction risk. The observations are n i.i.d. pairs $(X_i, Y_i) \in \mathcal{X} \times \{-1, +1\}$ and the aim is to find an estimator whose 0/1 risk is as close as possible to the minimum risk. The minimum risk is achieved by M^B, that is called the Bayes predictor and is defined by: $M^B_{i,j} = \operatorname{sign} \mathbb{E}(Y|X = E_{i,j})$.

Estimator and theoretical results.

The first step is to construct the estimator. We start by defining the empirical hinge risk:

$$\forall M \in \mathbb{R}^{m_1 \times m_2}, \quad r^h(M) = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - Y_i \langle M, X_i \rangle).$$

We use the prior distribution defined by (2.11) because it favors low rank matrices with an extra parameter. The parameter is finally $\theta = (L, R, \gamma)$. The pseudo posterior distribution is then defined in (2.10). It is very time-consuming for this problem to use a MCMC method so we propose a variational method in order to approximate the distribution. This approximation is similar to the one in (2.14) for a specific parametric family $\mathcal{F}^P$ that is well chosen. The bound

$$\forall f_m \in \mathcal{F}^P, \int \tau r^h(\theta) f_m(d\theta) + \mathcal{K}(f_m, \pi) = \mathcal{L}(m)$$

is not tractable with respect to the parameters m because the matrix M is factorized. We then use an upper bound:

$$\mathcal{L}(m) \leq AVB(m)$$

which is tractable. The distribution that is finally used is the distribution whose parameter minimizes this bound:$f_{\widehat{m}} : \widehat{m} = \min_{m \in \mathcal{M}} AVB(m)$.

The point estimate is thence defined by:

$$\widehat{M} = \int LR^\top f_{\widehat{m}}(d\theta).$$

We first study the theoretical properties of this estimator in the similar way of [Alquier et al., 2016]. We derive an empirical bound, in the sense that it is possible to compute it with the data. We also derive a theoretical bound; it needs a margin assumption, that is very common in classification task. These two bounds control the risk $R_{0/1}$ defined in (2.9). We remind here Theorem 2, that is the theoretical bound. We write $\overline{R} = \inf_M R_{0/1}(M)$.

Theorem 2.6 (Theorem 2 from Chapter 3)**.** *Assume that the margin assumption holds for $C > 0$. Then, for any $\varepsilon, s \in]0,1[$, with $\lambda = sn/C$, with probability greater than $1-\varepsilon$ we have:*

$$\int R_{0/1} df_{\widehat{m}} \leq 2(1+3s)\overline{R} + \mathcal{C}\left(\frac{\operatorname{rank}(M^B)\,\mathfrak{M}(\log n + L(M^B)) + \log 1/\varepsilon}{n}\right),$$

where L is a deterministic function and $\mathcal{C}$ is a numerical constant that depends on the prior distribution on γ.

In comparison to the previous results such as the one in Theorem 2.4, this result allows to compare the integrated 0/1 risk of the estimator to the best risk. In the noiseless case, the minimax rate of convergence is achieved up to a logarithm term. In the presence of noise, the result is quite limited because of the factor 2 that does not lead to the consistency of the estimator.

In practice.

The quantity $AVB(m)$ is not convex with respect to m because of the factorization of $M = LR^\top$. Nonetheless, it is biconvex in the sense that it is convex with respect to L and R. We then propose a coordinatewise optimization algorithm where, for the coordinates that have no explicit minimum, the step is replaced by a subgradient descent.

We can then test our estimator on simulated datasets in a first stage. Several scenarii are challenged: observations from a logistic model or not, a Bayes matrix that is low rank or not. For example, on Graph 2.2, we aim at recovering a rank 3 matrix. The performances of our estimator are better than the one from a GLM on medium level of noise. The

algorithms are then tested on the MovieLens dataset and it shows that our procedure suits large dataset.

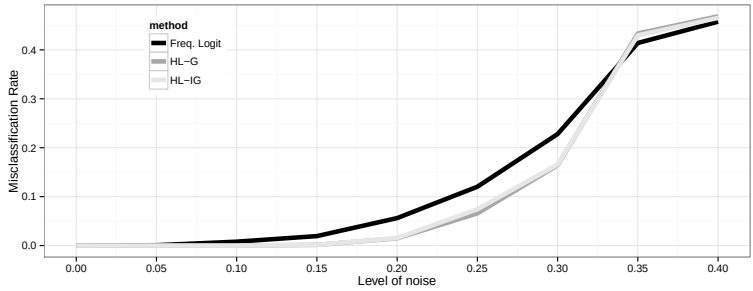


Figure 2.2 – 0/1 risk for the recovering of a rank 3 matrix (dimensions: 200×200) as a function of the level of noise. It is the *switch* noise (with probability p the observed value is the opposite to the true one).

Variational approximation for the logistic model.

In this chapter we also develop the variational approximation for the logistic model. The classic matrix completion problem is treated with this method in [Lim and Teh, 2007]. We need for the logistic model a double approximation. We also derive the algorithms for different prior distributions for γ. The variational approximations then use the *Generalized Inverse Gaussian* family.

2.4.2 Chapter 4: *Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions*

This chapter keeps exploring the matrix completion issue, but use a more general approach that allows to address other problems. It focuses on the properties of the regularized empirical risk minimizer when the loss function is Lipschitz. If we write $\ell(f(x), y)$ the loss occurred for a prediction by f at the point (x, y), this property is expressed by:

$$\forall (x,y), \forall (f_1, f_2), \quad |\ell(f_1(x), y) - \ell(f_2(x), y)| \leq |f_1(x) - f_2(x)| \,.$$

The Lipschitz property is verified for the following loss functions:

— loss for the quantile regression of order $\tau \in (0,1)$: $\ell(y', y) = \tau \max(0, y - y') + (1 - \tau) \max(0, y' - y)$ for any $y', y \in \mathbb{R}$
— logistic loss: $\ell(y', y) = \log(1+\exp(-yy'))$ for any $(y, y') \in \{-1, +1\} \times \mathbb{R}$
— hinge loss: $\ell(y', y) = \max(0, 1-yy')$ for any $(y, y') \in \{-1, +1\} \times \mathbb{R}$

Let $(X_i, Y_i)_{i=1}^n$ denote the observations. The space of predictors is written $\mathcal{F}$ and the norm over it is written $\|\cdot\|$. The estimator that will be considered is thus defined by:

$$\widehat{f} = \underset{f \in \mathcal{F}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \|f\| \right\}.$$

Although the regularized empirical risk minimizer has been well studied where the loss is the quadratic loss, there are less works about Lipschitz loss. Nevertheless, this property is very common in the classification framework (for $y \in \{-1, +1\}$) with the logistic and hinge loss. The quantile regression covers the ℓ_1 loss, that is a robust version of the quadratic loss; it may also treat the recovering of any quantile between 0 and 1.

Theoretical results.

The theoretical results are developed under two different assumptions that are made on X and $\mathcal{F}$: the subgaussian case and the bounded case. As these results are general, they depend on quantities that are computed with $(\mathcal{F}, \|\cdot\|)$. We write the oracle $f^\star$ and the excess risk of $f \in \mathcal{F}$ by $\mathcal{E}(f)$ where:

$$f^\star = \underset{f \in \mathcal{F}}{\arg\min}\, \mathbb{E}[\ell(f(X), Y)], \quad \mathcal{E}(f) = \mathbb{E}[\ell(f(X), Y)] - \mathbb{E}[\ell(f^\star(X), Y)].$$

We sketch here a summary of the strategy in order to effectively compute the non asymptotic bounds. The formal definitions of the quantities are in the chapter.

1. Compute the Bernstein parameter (κ, A) associated to the loss ℓ and the family $\mathcal{F}$.
2. Compute the complexity function

$$r(\rho) = \left[\frac{A\rho \mathrm{comp}(B)}{\sqrt{n}} \right]^{1/2\kappa},$$

where $\mathrm{comp}(B)$ depends on the framework (subgaussian or bounded) and the unit ball $B = \{f : \|f\| \leq 1\}$.

3. After computing the subdifferential of the norm $\partial \|\cdot\| (f^\star)$, we need to solve the sparsity equation. The potential parsimony of the oracle may help at this stage. $\rho^\star$ is then the radius satisfying:

$$\Delta(\rho^\star) = 4\rho^\star/5.$$

4. We are now ready to state the theorems with respect to the framework. With high probability, we have:

$$\left\| \widehat{f} - f^\star \right\| \leq \rho^\star, \quad (\mathbb{E}[\widehat{f}(X) - f^\star(X)]^2)^{1/2} \leq r(2\rho^\star),$$
$$\mathcal{E}(\widehat{f}) \leq C[r(2\rho^\star)]^{2\kappa}.$$

These results are obtained without any assumptions over Y, which shows the robustness of the Lipschitz loss functions. We may now apply the theoretical results to two problems: the logistic regression and the matrix completion. In the chapter, another example about RKHS is also developed.

Application 1: the logistic regression.

The known results about penalized logistic regression usually use the ℓ_1 norm on vectors. The SLOPE norm, introduced by [Bogdan et al., 2015], leads to the minimax rate for the least square regression but has not been used with the logistic loss. We define it, for any $t \in \mathbb{R}^p$ by

$$\|t\|_{SLOPE} = \sum_{j=1}^{p} \sqrt{\log(ep/j)} t_{(j)},$$

where the $t_{(j)}$'s are a non increasing rearrangement of $(|t_j|)$. The observations (X_i, Y_i) lie in $\{-1, +1\} \times \mathbb{R}^p$ and the estimator is therefore defined by:

$$\widehat{t} = \underset{t \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-Y_i \langle t, X_i \rangle)) + \frac{c}{\sqrt{n}} \|t\|_{SLOPE} \right\}.$$

In order to get the upper bound, we compute the quantities with respect to the variables. In the subgaussian framework, comp(B) is of order a constant. The Bernstein parameter κ is 1. As the subdifferential of the

SLOPE norm is wide when some coordinates are null, the sparsity equation gives $\rho^\star = c\frac{s}{\sqrt{n}}\log\frac{ep}{s}$ where s is the number of non null coordinates of $t^\star$. At the end, with high probability, we get:

$$\mathcal{E}_{\text{logistic}}(\widehat{t}) \leq c\frac{s\log ep/s}{n}$$

This result upgrades the known results using the ℓ_1 norm. In this case, the excess risk is controlled by a bound of order $s\log p/n$. When s is rather large, such that a fraction of p, the result with the SLOPE estimator is better. A same analysis can be made for the quantile regression loss (only the Bernstein parameter may change).

Application 2: matrix completion.

This part illustrates the bounded framework and leads to new results about matrix completion. The predictors are the $m_1 \times m_2$ matrices M whose entries are absolutely bounded by $\mathbf{a}$. The variables X_i lie in $\mathcal{X}$, the set of mask matrices. The regularization norm that will be used is the nuclear norm. The results that are presented here are only written for the Bernstein parameter $\kappa = 1$, that leads to fast rates. The general results are in the chapter. Hence we consider the following estimator:

$$\widehat{M} = \underset{M:\|M\|_\infty \leq \mathbf{a}}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(\langle M, X_i\rangle, Y_i) + \lambda\|M\|_{S_1}\right\}. \tag{2.16}$$

The complexity parameter is:

$$r(\rho) = c\left[\rho\sqrt{\frac{\log(m_1+m_2)}{n\,\mathfrak{m}}}\right]^{1/2}.$$

The nuclear norm has a wide subdifferential when the matrix is low rank. If the rank of $M^\star$ is s, the sparsity equation is verified by:

$$\rho^\star = csm_1m_2\left(\frac{\log(m_1+m_2)}{n\,\mathfrak{m}}\right)^{\frac{1}{2}}$$

Finally, with high probability, $\widehat{M}$ that satisfies (2.16) is such that:

$$\frac{1}{m_1m_2}\left\|\widehat{M} - M^\star\right\|_{S_2}^2 \leq c\frac{s\,\mathfrak{M}\log(m_1+m_2)}{n}$$

$$\mathcal{E}(\widehat{M}) \leq c\frac{s\,\mathfrak{M}\log(m_1+m_2)}{n}$$

This result is interesting in two ways: it bounds the reconstruction of $M^\star$ as it is often sought in matrix completion, but it also controls the excess risk. It is then possible to apply to some precise Lipschitz loss functions.

Binary matrix completion.

In binary matrix completion (when Y_i lies in $\{-1,+1\}$), one may use the logistic loss or the hinge loss. With the logistic loss, κ is 1 and we then recover the results of [Klopp et al., 2015] for the reconstruction of $M^\star$, but we also get a bound of the excess logistic risk of $\widehat{M}$. With the theorem from [Zhang, 2004], the 0/1 risk of $\widehat{M}$ is therefore upper bounded with high probability by $\sqrt{s\,\mathfrak{M}\log(m_1+m_2)/n}$.

It is then worth considering the hinge loss. The Bernstein parameter is 1 by assuming that $M^\star$ has no entries too close to 0. With this assumption, the estimator from (2.16) is such that, with high probability, the excess hinge risk is upper bounded by $s\,\mathfrak{M}\log(m_1+m_2)/n$. The 0/1 excess risk is then bounded by the same quantity, which is better than the previous known results.

The chapter also present lower bounds and a set of simulations. We propose an algorithm in order to compute the estimator $\widehat{M}$ which is of type *Alternating Direction Method of Multipliers* and is very similar to Algorithm 2.1. We test it on the MovieLens data and the notebooks are available online (http://sites.google.com/site/vincentcottet/code).

Matrix Completion using the quantile regression loss.

This loss function is rarely used in matrix completion although it can be employed for different tasks. For $\tau = 1/2$, the loss corresponds to the absolute value: it is a much more robust loss function to outliers than the quadratic loss. For other values of τ, it aims at recovering the quantiles exactly in a similar way to the quantile regression.

The Bernstein parameter κ is 1 as soon as the noise has a density with respect to the Lebesgue measure which is lower bounded in a enough large interval (it is easily computable for the classic distributions such as Gaussian, Student and even Cauchy). The results are then directly derived by applying the main theorem in the bounded case. For $\tau = 1/2$, if the noise is centered, the target matrix is the same as the one for the

quadratic loss. The reconstruction under the Frobenius norm is direct and the rate is the same as in [Klopp, 2014] or in [Mai and Alquier, 2015] and is optimal up to a logarithm term. The main advantage is that the assumptions on the noise are much wider (for the quadratic loss, the noise has to be subexponential).

We then test the robustness property with several scenarii of simulations. One of them involves outliers. We start with a rank 3 matrix of size 200×200 and 10% of the entries are corrupted by outliers. We then increase the magnitude of the outliers. The performance of the estimator from the quadratic loss is getting worse and worse and for a large magnitude, the estimated matrix is everywhere null. Conversely, the median estimator ($\tau = 1/2$) keeps almost the same performance even for a large magnitude of outliers, Figure 2.3. In the same spirit, the proportion of the outliers has a very small impact on the performance of this estimator (the details are in the chapter).

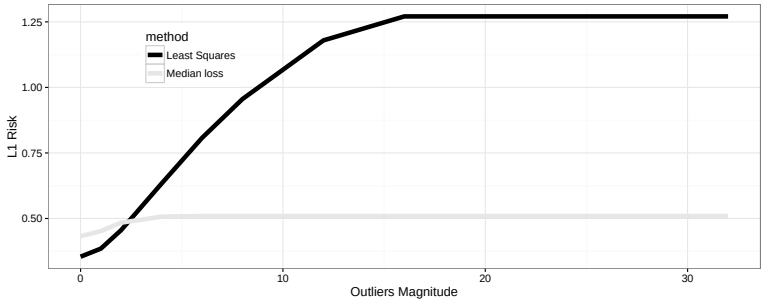


Figure 2.3 – Reconstruction in ℓ_1 norm of a low rank matrix. The magnitude of the outliers increases from 0 to 30.

2.4.3 Chapter 5: *Divide and Conquer in ABC: Expectation-Propagation algorithms for likelihood-free inference*

The two former chapters deal with high dimensional problems: the parameter is very large and we seek a parsimonious one that fits the data. The ABC methods, *Approximate Bayesian Computation*, have been developed thanks to the massive increasing of the computer capabilities and allow to tackle new problems. They aim at doing Bayesian inference for models where the likelihood is either not explicit or too expensive to compute. For these models, the parameter is very hard to estimate.

All what we need for ABC is that we are able to simulate data from the likelihood: the evaluation of it is not required. These models are common in biology, neuroscience or for spatial models.

The chapter is an extension of [Barthelmé and Chopin, 2014]. We first introduce the ABC algorithm and then present the EP approximation in order to conclude with the EP-ABC algorithm. At the end the illustrative example is exposed.

The ABC algorithm is very simple and seems very natural. We have at hand a prior distribution $p(\theta)$, a likelihood $p(x|\theta)$ and observations $x^\star$ from the model. We need a distance between the observed and simulated data $d(x^\star, x)$ and a threshold $\varepsilon > 0$. The algorithm is then to repeat the following steps until we get enough accepted values:

1. draw $\theta \sim p(\theta)$
2. draw a sample of data $x \sim p(x|\theta)$
3. accept θ if $d(x^\star, x) \leq \varepsilon$.

Eventually, the accepted values is a sample of the following distribution:

$$p_\varepsilon(\theta|x^\star) \propto p(\theta) \int p(x|\theta) \mathbb{1}_{\{d(x,x^\star)\leq\varepsilon\}} dx.$$

An important limit of this method is the choice of the distance d. If it is possible to take $d(x^\star, x) = \|s(x^\star) - s(x)\|$ where s is a sufficient statistic and $\|\cdot\|$ a norm, then the approximation will tend to the true posterior distribution if ε tends to 0. If it is not the case, it is a double approximation that is hard to control.

Divide and Conquer: split the problem in blocks.

Many times the data may be gathered in a natural way into blocks, written $(x_i)_{i=1}^k$; each block may be of different size. We write $x_{[k]} = (x_1, \ldots, x_k)$. The likelihood may be written sequentially by block [6]:

$$p(x|\theta) = \prod_{i=1}^{k} p(x_i|x_{[i-1]}, \theta).$$

If it is possible to draw x_i from $p(x_i|x_{[i-1]}, \theta)$, we can therefore split the problem into blocks and approximate the distribution:

$$p_\varepsilon(\theta|x^\star) \propto p(\theta) \prod_{i=1}^{k} \int p(x_i|x^\star_{[i-1]}, \theta) \mathbb{1}_{\{\|s_i(x_i) - s_i(x_i^\star)\| \leq \varepsilon\}} dx_i, \qquad (2.17)$$

6. This writing is always correct and needs no assumption.

where s_i is a statistic of x_i. The EP-ABC algorithm aims at making an EP-type approximation of the distribution $p_\varepsilon(\theta|x^\star)$.

The first case in which the likelihood is naturally split is when the data are i.i.d. or i.i.d. by blocks. For example, if we observe several years of a geographical process such the one presented later, we can assume that the observations by year are mutually independent. A second case arises when the data are time series.

Expectation Propagation: A Bayesian approximation.

The EP algorithm, *Expectation Propagation*, is quite similar to the VB approach: the goal is to find a distribution q in a particular parametric family $\mathcal{Q}$ that will be close to the target distribution π. It is used when π is factorized into n factors such that:

$$\pi(\theta) = \frac{1}{Z_\pi} \prod_{i=0}^{k} l_i(\theta).$$

The idea is then to approximate each *site* l_i by a factor q_i and the final approximation will be

$$q(\theta) = \prod_{i=0}^{k} q_i(\theta).$$

To compute so, the algorithm looks like a coordinatewise optimization. Each site is updated sequentially while other sites are kept fixed. The update of the i-th site follows the steps:

1. Compute the *cavity distribution* $q_{-i}(\theta) \propto \prod_{j \neq i} q_j(\theta)$. The hybrid distribution is $h_i \propto q_{-i}(\theta) l_i(\theta)$.
2. Update q_i such as
$$q \in \arg\min_{f \in \mathcal{Q}} \mathcal{K}(h_i, f). \tag{2.18}$$

The last step is crucial and is explicit if $\mathcal{Q}$ is an exponential family. In this case, the moments of h_i and q has to be the same. The chapter uses Gaussian distributions as approximation; the main step (2.18) is translated into the computation of the moments:

$$\mu_h = \frac{1}{\int h_i(\theta) d\theta} \int \theta h_i(\theta) d\theta$$
$$\Sigma_h = \frac{1}{\int h_i(\theta) d\theta} \int \theta\theta^\top h_i(\theta) d\theta - \mu_h \mu_h^\top$$

In practice, the results are pretty good and the approximation seems to be better than the one from VB. The EP algorithm indeed keeps a full covariance for the parameter and thus between its coordinates. In opposition, the *mean-field* VB approximation will not conserve this flexibility and will approach it by independent components. Nonetheless, in a prediction approach, it is sometimes not required to have the covariance information and the VB approximation would be enough.

The EP algorithm suffers from a lack of theoretical works. Two recent articles [Dehaene and Barthelmé, 2015b] and [Dehaene and Barthelmé, 2015a] partially fill the gap: the first one studies the quality of the fixed point of the algorithm. The authors show that the mean and the variance are close to the ones from the target distribution for a Gaussian approximation. The second article focuses on the algorithm and shows that it may be interpreted as a Newton algorithm of optimization. This fact implies that the updates are not very stable and may face some difficulties to converge. It may be fixed by averaging the updates.

The EP-ABC algorithm in practice.

The target distribution defined in (1.15) is factorized into $k+1$ blocks (the first one being the prior distribution). Each site is:

$$l_i(\theta) = \int p(x_i|x^\star_{[i-1]}, \theta)\mathbb{1}_{\{\|s_i(x_i)-s_i(x^\star_i)\|\leq\varepsilon\}}dx_i.$$

The algorithm is conceivable if it is possible to compute the moments of the hybrid distributions. To do so, we use an ABC approximation, that is itself possible if we can simulate from the density $p(x_i|x^\star_{[i-1]})$. The full algorithm is given in the chapter.

What we propose has important advantages: at each step, some θs are drawn from the cavity distributions q_i that is usually more precise than the prior distribution [7]. Furthermore, the simulation of x_i and the associated acceptation of θ is a related to $\|s_i(x_i) - s_i(x^\star_i)\|$. As the dimension is smaller, the acceptance probability is larger, or it is possible to reduce the threshold. It avoids the curse of dimensionality: for ABC, adding data leads to a lower acceptance rate and a deterioration of the approximation quality [8]. With the EP-ABC, the splitting into blocks leads to a neutral addition of points if they are gathered in new blocks.

7. This point is often improved in advanced versions of ABC.

8. Here the dimensionality concerns the number of points, not the size of the parameter.

Moreover, a distributed version is easily derived. The most expensive step, that is the moment computation, may be done in a parallel way thanks to the splitting. This algorithm smooths the updates and is more stable.

An application to spatial extremes.

Three examples are treated in the seminal paper [Barthelmé and Chopin, 2014]: n i.i.d. observations from an alpha-stable distribution (distribution that has no explicit density), a Lotka-Volterra model (also known as predator-prey model) and a model about reaction time where time series are modeled jointly. In the last two examples, the sequential form is natural.

In Chapter 5, we propose another example where the data are k blocks of n points that are dependent into blocks [9]. The observations are the realizations of a spatial process observed in several locations written $(x_i)_{i=1}^n$. The observation for the j-th year at location x_i is $y_j(x_i)$ and we assume that the observations across year are independent. The process we consider here is a max-stable process; these process are used to model extrema. The goal is to infer the covariance of the process in order to get an idea of the dependence across location. For more than two locations, the likelihood is not explicit. A composite likelihood method has been proposed, that is an approximated frequentist method. It is asymptotically valid but it is not the considered case here.

The classic ABC method has been proposed for this model in [Erhardt and Smith, 2012]. It is very time-consuming because the simulation of a max-stable process is very demanding, but also the acceptance rate is very small. Nevertheless, this article offers many statistics and is a good starting point.

The EP-ABC is a large help in this case because of the independence between years. It allows to treat each block separately and we avoid an extra approximation. It also allows to use the recycling of the simulations, that is very helpful for expensive simulations. Finally, we apply this model to rainfall maxima in Switzerland. The maximum is computed every year from June to August between 1962 and 2008. We chose the Whittle-Matern covariance function with two parameters (c, ν).

The algorithm converges quite quickly, generally after 3 or 4 passes over the data. The time saving is huge in comparison to the classic

9. The number of points per block may vary but it is not the case here.

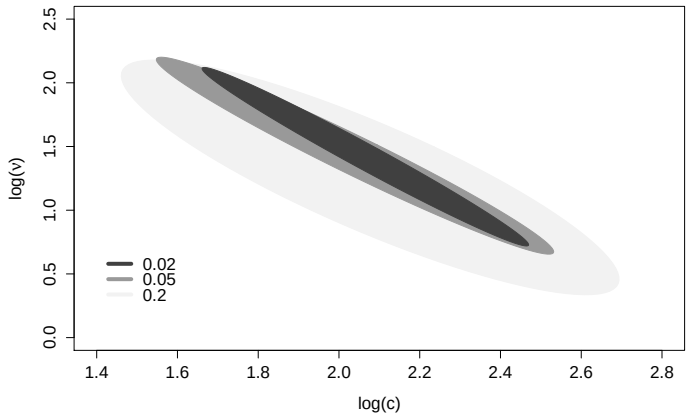


Figure 2.4 – 50% credible ellipses of the Gaussian approximation of the posterior distribution for different values of ε.

ABC. By testing different values of ε, we see on Graph 2.4 that the distribution of the posterior distribution is very correlated between the two parameters.

Chapter 3

1-bit Matrix Completion: PAC-Bayesian Analysis of a Variational Approximation

With Pierre Alquier. To appear in *Machine Learning Journal*.

Abstract

We focus on the completion of a (possibly) low-rank matrix with binary entries, the so-called 1-bit matrix completion problem. Our approach relies on tools from machine learning theory: empirical risk minimization and its convex relaxations. We propose an algorithm to compute a variational approximation of the pseudo-posterior. Thanks to the convex relaxation, the corresponding minimization problem is bi-convex, and thus the method works well in practice. We study the performance of this variational approximation through PAC-Bayesian learning bounds. Contrary to previous works that focused on upper bounds on the estimation error of M with various matrix norms, we are able to derive from this analysis a PAC bound on the prediction error of our algorithm.

We focus essentially on convex relaxation through the hinge loss, for which we present a complete analysis, a complete simulation study and a test on the MovieLens data set. We also discuss a variational approximation to deal with the logistic loss.

3.1 Introduction

Motivated by modern applications like recommendation systems and collaborative filtering, video analysis or quantum statistics, the matrix completion problem has been widely studied over the recent years. Recovering a matrix is, without any additional information, a impossible task. However, under some assumptions on the structure of the matrix to be recovered, it might become feasible, as shown by Candès and Tao [2010] and Candès and Recht [2012] where the assumption is that the matrix has a small rank. This assumption is natural in many applications. For example, in recommendation systems, it is equivalent to the existence of a small number of hidden features that explain the users preferences. While [Candès and Tao, 2010] and [Candès and Recht, 2012] focused on matrix completion without noise, many authors extended these techniques to the case of noisy observations, see [Candès and Plan, 2010] and [Chatterjee, 2015] among others. The main idea in [Candès and Plan, 2010] is to minimize the least squares criterion, penalized by the rank. This penalization is then relaxed by the nuclear norm, which is the sum of the singular values of the matrix at hand. An efficient algorithm is described in [Recht and Ré, 2013].

All the aforementioned papers focused on real-valued matrices. However, in many applications, the matrix entries are binary, that is in the set $\{0, 1\}$. For example, in collaborative filtering, we have often only access to a binary choice: the $(i, j) - th$ entry being 1 means that user i is satisfied by object j while this entry being 0 means that he/she is not satisfied by it. The problem of recovering a binary matrix from partial observations is usually referred as 1-bit matrix completion. To deal with binary observations requires specific estimation methods. Most works on this problem usually assume a generalized linear model (GLM): the observations Y_{ij} for $1 \leq i \leq m_1$, $1 \leq j \leq m_2$, are Bernoulli distributed with parameter $f(M_{ij})$, where f is a link function which maps from $\mathbb{R}$ to $[0, 1]$, for example the logistic function $f(x) = \exp(x)/[1 + \exp(x)]$, and M is a $m_1 \times m_2$ real matrix, see [Cai and Zhou, 2013, Davenport et al., 2014, Klopp et al., 2015]. In these works, the goal is to recover the matrix M and a convergence rate is then derived. For example, [Klopp et al., 2015] provides an estimate $\widehat{M}$ for which, under suitable assumptions and when

the data are generated according to the true model with $M = M_0$,

$$\frac{1}{m_1 m_2}\|\widehat{M} - M_0\|_F^2 \leq C \max\left(\sqrt{\frac{\log(m_1+m_2)}{n}}, \frac{\max(m_1,m_2)\mathrm{rank(M_0)}\log(m_1+m_2)}{n}\right)$$

for some constant C that depends on the assumptions and the sampling scheme, and where $\|.\|_F$ stands for the Frobenius norm (we refer the reader to Corollary 2 page 2955 in [Klopp et al., 2015] for the exact statement). While this result ensures the consistency of $\widehat{M}$ when M_0 is low-rank, it does not provide any guarantee on the probability of a prediction error. Moreover, the results rely on the assumption that the model (in particular the function f) is well specified. In practice, this assumption is unrealistic, and it is important to provide generalization error bounds that hold even in case of misspecification.

Here, we adopt a machine learning point of view: in machine learning, dealing with binary output is called a classification problem, for which methods are known that do not assume any model on the observations. That is, instead of focusing on a parametric model for $Y_{i,j}$, we will only define a set of prediction matrices M and seek for the one that leads to the best prediction error. Using the zero-one loss function, we could actually directly use Vapnik-Chervonenkis theory [Vapnik, 1998] to propose a classifier $\widehat{M}$ risk would be controlled by a PAC inequality. However, it is known that this approach usually is computationally intractable. A popular approach is to replace the zero-one loss by a convex surrogate [Zhang, 2004], namely, the hinge loss. Our approach is as follows: we propose a pseudo-Bayesian approach, where we define a pseudo-posterior distribution on a set of matrices M. This pseudo-posterior distribution does not have a simple form, however, thanks to a variational approximation, we manage to approximate it by a tractable distribution. Thanks to the PAC-Bayesian theory [McAllester, 1998, Herbrich and Graepel, 2002, Shawe-Taylor and Langford, 2003, Catoni, 2004, 2007, Seldin et al., 2012, Dalalyan and Tsybakov, 2008], we are able to provide a PAC bound on the prediction risk of this variational approximation. We then show that, due to the convex relaxation of the zero-one loss, the computation of this variational approximation is actually a bi-convex minimization problem. As a consequence, efficient algorithms are available.

Other settings for 1-bit matrix completion have also been studied. For example, in some real-life applications, only positive instances are

available. This setting is studied in details in [Hsieh et al., 2015]. It requires a different approach. Here, we stick to the classification approach where positive and negative instances are observed. We refer the reader to [Hsieh et al., 2015] and the references therein for the positive-only case.

The rest of the paper is as follows. In Section 3.2 we provide the notations used in the paper, the definition of the pseudo-posterior and of its variational approximation. In Section 3.3 we give the PAC analysis of the variational approximation. This yields an empirical and a theoretical upper bound on the prediction risk of our method. Section 3.4 provides details on the implementation of our method. Note that in the aforementioned sections, the convex surrogate of the zero-one loss used is the hinge loss. An extension to the logistic loss is briefly discussed in Section 3.5, together with an algorithm to compute the variational approximation. Finally, Section 3.6 is devoted to an empirical study and Section 3.6.3 to an application to the MovieLens data set. The proof of the theorems of Section 3.3 are provided in Section 3.8.

3.2 Estimation Procedure

For any integer m we define $[m] = \{1, \dots, m\}$; for two real numbers a and b we write $\max(a, b) = a \vee b$ and $\min(a, b) = a \wedge b$. We define, for any integers m_1 and m_2 and any matrix $M \in \mathbb{R}^{m_1 \times m_2}$, $\|M\|_{\max} = \max_{(i,j) \in [m_1] \times [m_2]} M_{ij}$. Let $\mathbb{R}^+$ stand for the set of non-negative real numbers, and $\mathbb{R}^{+*}$ for the positive real numbers. For any real number a, $(a)_+$ is the positive part of a and is equal to $\max(0, a)$.

For a pair of matrices (A, B), we write $\ell(A, B) = \|A\|_{\max} \vee \|B\|_{\max}$. Finally, when an $m_1 \times m_2$ matrix M has $\operatorname{rank}(M) = r$ then it can be written as $M = LR^T$ where L is $m_1 \times r$ and R is $m_2 \times r$. This decomposition is obviously not unique; we put $\ell(M) = \inf_{(L,R)} \ell(L, R)$ where the infimum is taken over all such possible pairs (L, R) such that $LR^\top = M$. In frequentist approaches like [Klopp et al., 2015], it is common that the upper bound depends on the infinite norm of the entries. This quantity is replaced in our analysis by $\ell(M)$.

3.2.1 1-bit matrix completion as a classification problem

We formally describe the 1-bit matrix completion problem as a classification problem: we observe $(X_k, Y_k)_{k \in [n]}$ that are n i.i.d pairs from a distribution $\mathbf{P}$. The X_k's take values in $\mathcal{X} = [m_1] \times [m_2]$ and the Y_k's take

values in $\mathcal{Y} = \{-1, +1\}$. Hence, the k-th observation of an entry of the matrix is Y_k and the corresponding position in the matrix is provided by $X_k = (i_k, j_k)$. In this setting, a predictor is a function $[m_1] \times [m_2] \to \mathbb{R}$ and thus can be represented by a matrix M and for any X, M_X is the entry of M at location X. It is natural to use M in the following way: when $(X, Y) \sim \mathbf{P}$, M predicts Y by $\text{sign}(M_X)$. The ability of this predictor to predict a new entry of the matrix is then assessed by the risk

$$\mathbf{R}(M) = \mathbb{E}_{\mathbf{P}} \left[\mathbb{1}(Y M_X < 0) \right],$$

and its empirical counterpart is:

$$r_n(M) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}(Y_k M_{X_k} < 0) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}(Y_k M_{i_k, j_k} < 0).$$

It is then possible to use the standard approach in classification theory [Vapnik, 1998]. For example, the best possible classifier is the Bayes classifier and it relies on the regression function:

$$\eta(x) = \mathbb{E}(Y|X = x) \quad \text{or equivalently} \quad \eta(i, j) = \mathbb{E}[Y|X = (i, j)],$$

and therefore we have an optimal matrix

$$M^B_{ij} = \text{sign}[\eta(i, j)] = \text{sign}\Big\{ \mathbb{E}[Y|X = (i, j)] \Big\}.$$

We define $\overline{\mathbf{R}} = \inf_M \mathbf{R}(M) = \mathbf{R}(M^B)$, and $\overline{r_n} = r_n(M^B)$. Note that, clearly, if two matrices M^1 and M^2 are such as, for every (i, j), $\text{sign}(M^1_{ij}) = \text{sign}(M^2_{ij})$ then $\mathbf{R}(M^1) = \mathbf{R}(M^2)$, and obviously,

$$\forall M, \forall (i, j) \in [m_1] \times [m_2], \quad \text{sign}(M_{ij}) = M^B_{ij} \quad \Rightarrow \quad r_n(M) = \overline{r_n}.$$

While the risk $\mathbf{R}(M)$ has a clear interpretation, its empirical counterpart $r_n(M)$ usually leads to intractable problems, as it is non-smooth and non-convex. Hence, it is standard to replace the empirical risk by a convex surrogate [Zhang, 2004]. In this paper, we will mainly deal with the hinge loss, which leads to the following so-called hinge risk and hinge empirical risk:

$$R^h(M) = \mathbb{E}_{\mathbf{P}} \left[(1 - Y M_X)_+ \right],$$
$$r^h_n(M) = \frac{1}{n} \sum_{k=1}^{n} (1 - Y_k M_{X_k})_+.$$

The hinge loss was also used by Srebro et al. [2005] and in [Herbster et al., 2016] in the 1-bit matrix completion problem, with a different approach leading to different algorithms. Moreover, here, we provide an analysis of the rate of convergence of our method, that is not provided in [Srebro et al., 2005]. Note that our analysis can be extended to any Lipschitz and convex surrogate, and indeed, we study also briefly the logistic loss in Section 3.5. Still, we prefer to focus only on the hinge loss in the main part of the paper, for its good algorithmic and theoretical properties, c.f. [Zhang, 2004].

Contrary to many recent papers on matrix completion, our approach leads to distribution-free bounds. The marginal distribution of X is not an issue and we do not have to assume a uniform sampling scheme. Following standard notations in matrix completion, we define Ω as the set of indices of observed entries: $\Omega = \{X_1, \dots, X_n\}$. We will use in the following the sub-sample of $\{1, \dots, n\}$ for a specified line i: $\Omega_{i,\cdot} = \{l \in [n] : (i, j_l) \in \Omega\}$ and the counterpart for a specified column j: $\Omega_{\cdot,j} = \{l \in [n] : (i_l, j) \in \Omega\}$.

3.2.2 Pseudo-Bayesian estimation

The Bayesian framework has been used several times for matrix completion (see [Salakhutdinov and Mnih, 2008, Lim and Teh, 2007] and the references therein). The PAC-Bayesian approach has been well used on different models (see [Mai and Alquier, 2015] and Section 6 in [Seldin and Tishby, 2010]). A common idea in all of these papers is to factorize the matrix into two parts in order to define a prior on low-rank matrices. It needs an additional parameter and a hierarchical model in order to be rank-adaptive and we explain here the idea. Every matrix whose rank is r can be factorized:

$$M = LR^\top, L \in \mathbb{R}^{m_1 \times r}, \quad R \in \mathbb{R}^{m_2 \times r}.$$

As mentioned in the introduction, the Bayes matrix M^B is expected to be low-rank, or at least well approximated by a low-rank matrix. However, in practice, we do not know what would be the rank of this matrix. So, we actually write $M = LR^\top$ with $L \in \mathbb{R}^{m_1 \times K}$, $R \in \mathbb{R}^{m_2 \times K}$ for some large enough K. Adaptation with respect to $r \in [K]$ is obtained by shrinking some columns of L and R to 0. In order to do so, we will scale parameters γ_k for the columns of L and R, and let $\gamma := (\gamma_1, \dots, \gamma_K)$. We then define

the following hierarchical probability distribution:

$$\forall k \in [K], \quad \gamma_k \overset{iid}{\sim} \pi^\gamma, \tag{3.1}$$

$$\forall (i,j,k) \in [m_1] \times [m_2] \times [K], \quad L_{i,k}, R_{j,k} | \gamma \overset{indep.}{\sim} \mathcal{N}(0, \gamma_k), \tag{3.2}$$

$$\text{and } M = LR^\top, \tag{3.3}$$

where the prior distribution on the variances π^γ is yet to be specified. It means that the entries of L and R are normally distributed but the variance depends on the column index: a large γ_k leads to spread values and a small γ_k leads to almost null entries of the column k. In most papers π^γ is chosen as an inverse-Gamma distribution because it is conjugate in this model. This kind of hierarchical prior distribution is also very similar to the Bayesian Lasso developed in [Park and Casella, 2008] and especially of the form of the Bayesian Group Lasso developed in [Kyung et al., 2010] in which the variance term is Gamma distributed. We will show that the Gamma distribution is a possible alternative in matrix completion, both for theoretical results and practical considerations. Thus all the results in this paper are stated under the assumption that π^γ is either the Gamma or the inverse-Gamma distribution: $\pi^\gamma = \Gamma(\alpha, \beta)$, or $\pi^\gamma = \Gamma^{-1}(\alpha, \beta)$.

Let θ denote the parameter $\theta = (L, R, \gamma)$ and π denote the prior distribution defined in (3.1). Following the aforementioned papers in PAC-Bayesian theory, we define the pseudo-posterior as follows:

$$\widehat{\rho}_\lambda(d\theta) = \frac{\exp[-\lambda r_n^h(LR^\top)]}{\int \exp[-\lambda r_n^h] d\pi} \pi(d\theta)$$

where $\lambda > 0$ is a parameter to be fixed by the statistician. The calibration of λ is discussed below. This distribution is close to a classic posterior distribution but the likelihood has been replaced by the pseudo-likelihood $\exp[-\lambda r_n^h(LR^\top)]$ based on the hinge empirical risk.

3.2.3 Variational Bayes approximations

Unfortunately, the pseudo-posterior is intractable and MCMC methods may be too expensive because of the dimension of the parameter. We decide to use a Variational Bayes approximation, that is to seek an approximation of the pseudo-posterior by efficient optimization algorithms [Bishop, 2006]. First, we fix a subset $\mathcal{F}$ of the set of all distributions on the parameter space. The class $\mathcal{F}$ should be large enough to contain a good enough approximation of $\widehat{\rho}_\lambda$, but not too large, in order to

keep optimization in $\mathcal{F}$ feasible. The VB approximation is then defined by

$$\arg\min_{\rho\in\mathcal{F}} \mathcal{K}(\rho, \widehat{\rho}_\lambda), \text{ where } \mathcal{K}(\rho, \widehat{\rho}_\lambda) = \int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\widehat{\rho}_\lambda}\right) \mathrm{d}\rho$$

is the Kullback-Leibler divergence between ρ and $\widehat{\rho}_\lambda$. When $\mathcal{K}(\rho, \widehat{\rho}_\lambda)$ is not available in closed form, it is usual to replace it by an upper bound. Following the classical approach with matrix factorization priors, as in [Lim and Teh, 2007], we define here the class $\mathcal{F}$ as follows:

$$\mathcal{F} = \Bigg\{ \rho(\mathrm{d}L, \mathrm{d}R, \mathrm{d}\gamma) = \prod_{k=1}^{K} \Bigg[\prod_{i=1}^{m_1} \varphi(L_{i,k}; L^0_{i,k}, v^L_{i,k}) \mathrm{d}L_{i,k} \prod_{j=1}^{m_2} \varphi(R_{j,k}; R^0_{j,k}, v^R_{j,k}) \mathrm{d}R_{j,k} \rho^{\gamma_k}(d\gamma_k) \Bigg], \; L^0 \in \mathbb{R}^{m_1\times K}, R^0 \in \mathbb{R}^{m_2\times K}, v^L \in \mathbb{R}_+^{m_1\times K}, v^R \in \mathbb{R}_+^{m_2\times K} \Bigg\},$$

where $\varphi(.; \mu, v)$ is the density of the Gaussian distribution with parameters (μ, v) and ρ^{γ_k} ranges over all possible probability distributions for $\gamma_k \in \mathbb{R}^+$. The VB approximations are referred as *parametric* when $\mathcal{F}$ is finite dimensional and as *mean-field* otherwise. Here we actually use a mixed approach. Informally, under $\rho \in \mathcal{F}$, all the coordinates are independent and the variational distribution of the entries of L and R is specified. The free variational parameters to be optimized are the means and the variances. We will show below that the optimization with respect to ρ^{γ_k} is available in close form. Also, note that any probability distribution $\rho \in \mathcal{F}$ is uniquely determined by L^0, R^0, v^L, v^R and $\rho^{\gamma_1}, \ldots, \rho^{\gamma_K}$. We could actually use the notation $\rho = \rho_{L^0,R^0,v^L,v^R,\rho^{\gamma_1},\ldots,\rho^{\gamma_K}}$, but it would be too cumbersome, so we will avoid it as much as possible. Conversely, once ρ is given in $\mathcal{F}$, we can define $L^0 = \mathbb{E}_\rho[L]$, $R^0 = \mathbb{E}_\rho[R]$ and so on.

It is well-known that the Kullback divergence can be decomposed as

$$\mathcal{K}(\rho, \widehat{\rho}_\lambda) = \lambda \int r_n^h d\rho + \mathcal{K}(\rho, \pi) + \log \int \exp[-\lambda r_n^h] d\pi \qquad (3.4)$$

but the first term in the right-hand side is not tractable here. We then use the Lipschitz property of the loss and derive an upper bound of the Kullback divergence for any $\rho \in \mathcal{F}$ which is explicit in the parameters of ρ. It is this quantity that we will optimize in the algorithm and we

will see in the next section that this estimate enjoys good properties. We remind the reader that all the proofs are postponed to Section 3.8.

Proposition 3.1. *For any* $\rho = \rho_{L^0,R^0,v^L,v^R,\rho^{\gamma_1},\ldots,\rho^{\gamma_k}} \in \mathcal{F}$,

$$\int r_n^h d\rho + \frac{1}{\lambda}\mathcal{K}(\rho,\pi) \leq r_n^h\left(\mathbb{E}_\rho[L]\mathbb{E}_\rho[R]^\top\right) + \mathcal{R}(\rho,\lambda) \tag{3.5}$$

where

$$\begin{aligned}
\mathcal{R}(\rho,\lambda) = \frac{1}{n}\sum_{h=1}^{n}\sum_{k=1}^{K}&\left[\sqrt{v^L_{i_h,k}\frac{2}{\pi}}\sqrt{v^R_{j_h,k}\frac{2}{\pi}} + |R^0_{j_h,k}|\sqrt{v^L_{i_h,k}\frac{2}{\pi}} + |L^0_{i_h,k}|\sqrt{v^R_{j_h,k}\frac{2}{\pi}}\right]\\
&+\frac{1}{\lambda}\left\{\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_\rho\left[\frac{1}{\gamma_k}\right]\left(\sum_{i=1}^{m_1}(v^L_{i,k}+(L^0_{i,k})^2)+\sum_{j=1}^{m_2}(v^R_{j,k}+(R^0_{j,k})^2)\right)\right.\\
&-\frac{1}{2}\sum_{k=1}^{K}\left(\sum_{i=1}^{m_1}\log v^L_{i,k}+\sum_{j=1}^{m_2}\log v^R_{j,k}\right)\\
&\left.+\sum_{k=1}^{K}\left[\mathcal{K}(\rho^{\gamma_k},\pi^\gamma)+\frac{m_1+m_2}{2}\left(\mathbb{E}_\rho\left[\log\gamma_k\right]-1\right)\right]\right\}.
\end{aligned}$$

Note that the explicit expression of our upper bound $\mathcal{R}(\rho,\lambda)$ is very cumbersome to say the least. A few comments are in order. First, this upper bound is explicit and can be computed easily. Hence, instead of minimizing the Kullback divergence, this is this term that we minimize in practice. Then, our theoretical analysis will show that the upper bound is acceptable in the sense that its minimization will lead to a small generalization error. But it is actually possible to understand at first sight why the minimization of $r_n^h\left(\mathbb{E}_\rho[L]\mathbb{E}_\rho[R]^\top\right)+\mathcal{R}(\rho,\lambda)$ works well. Indeed, assume that a matrix M with $r_n^h(M)=0$ satisfies $\mathrm{rank}(M)=r\ll K$. Then, it is possible to decompose M as a product $M=L^0(R^0)^\top$ with $L^0_{i,k}=R^0_{j,k}=0$ when $r<k\leq K$. So, the sum

$$\frac{1}{2\lambda}\sum_{k=1}^{K}\mathbb{E}_\rho\left[\frac{1}{\gamma_k}\right]\left(\sum_{i=1}^{m_1}(L^0_{i,k})^2+\sum_{j=1}^{m_2}(R^0_{j,k})^2\right)$$

has actually only $r(m_1+m_2)=\mathrm{rank}(M)(m_1+m_2)$ non-null terms. To minimize $r_n^h\left(\mathbb{E}_\rho[L]\mathbb{E}_\rho[R]^\top\right)+\mathcal{R}(\rho,\lambda)$ is thus related to penalized risk minimization with a penalty proportional to the rank, as in most frequentist approaches [Klopp et al., 2015].

The quantity $r_n^h\left(\mathbb{E}_\rho[L]\mathbb{E}_\rho[R]^\top\right)+\mathcal{R}(\rho,\lambda)$ will be referred as the Approximate Variational Bound (AVB) of ρ in the following. We are now able to define our estimate.

Definition 3.1. *For a fixed $\lambda>0$ we put*

$$\begin{aligned} AVB(\rho,\lambda) &= r_n^h\left(\mathbb{E}_\rho[L]\mathbb{E}_\rho[R]^\top\right)+\mathcal{R}(\rho,\lambda),\\ \widetilde{\rho}_\lambda &= \arg\min_{\rho\in\mathcal{F}} AVB(\rho,\lambda). \end{aligned} \tag{3.6}$$

Also, when explicit notations involving $L^0,R^0,v^L,v^R,\rho^{\gamma_1},\ldots,\rho^{\gamma_k}$ are necessary we will use the notation

$$AVB(L^0,R^0,v^L,v^R,\rho^{\gamma_1},\ldots,\rho^{\gamma_k},\lambda)=AVB(\rho_{L^0,R^0,v^L,v^R,\rho^{\gamma_1},\ldots,\rho^{\gamma_k}},\lambda).$$

In the next section, we study the theoretical properties of our estimate. The main result is that the minimizer $\widetilde{\rho}_\lambda$ of the $AVB(\rho,\lambda)$ has a small prediction risk for a well chosen λ. We also provide an algorithm that computes $\widetilde{\rho}_\lambda$ and show on simulations that it behaves well in practice.

3.3 PAC analysis of the variational approximation

Paper [Alquier et al., 2016] proposes a general framework for analyzing the prediction properties of VB approximations of pseudo-posteriors based on PAC-Bayesian bounds. In this section, we apply this method to derive a control of the out-of-sample prediction risk $\mathbf{R}$ for our approximation $\widetilde{\rho}_\lambda$.

3.3.1 Empirical Bound

The first result is a so-called empirical bound: it provides an upper bound on the prediction risk of the pseudo-posterior $\widetilde{\rho}_\lambda$ that depends only on the data and on quantities defined by the statistician.

Lemma 3.1. *For any $\epsilon\in(0,1)$, with probability at least $1-\epsilon$ on the drawing of the sample, for any $\rho\in\mathcal{F}$,*

$$\int \mathbf{R}d\rho \leq r_n^h\left(\mathbb{E}_\rho[L]\mathbb{E}_\rho[R]^\top\right)+\mathcal{R}(\rho,\lambda)+\frac{\lambda}{2n}+\frac{\log\frac{1}{\epsilon}}{\lambda}=AVB(\rho,\lambda)+\frac{\lambda}{2n}+\frac{\log\frac{1}{\epsilon}}{\lambda}.$$

This shows that our strategy to minimize $AVB(\rho, \lambda)$ is indeed the minimization of an empirical upper bound on the prediction risk, a standard approach in PAC-Bayesian theory. An immediate consequence of Lemma 3.1 and of the definition of $\widetilde{\rho}_\lambda$ is the following theorem.

Theorem 3.1. *For any $\epsilon \in (0,1)$, with probability at least $1-\epsilon$ on the drawing of the sample,*

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \inf_{\rho \in \mathcal{F}} AVB(\rho, \lambda) + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda}$$

Even though the bound in the right-hand side may be evaluated in practice, and thus may provide a numerical guarantee on the out-of-sample prediction risk, it is not very clear how it depends on the parameters. The following corollary of Theorem 3.1 will clarify things. It is obtain by deriving upper bounds of $AVB(\rho, \lambda)$ (once again, the proof is provided explicitly in Section 3.8).

Corollary 3.1. *Assume that $\lambda \leq n$. For any $\epsilon \in (0,1)$,with probability at least $1-\epsilon$:*

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \inf_{M} \left[r_n^h(M) + \mathcal{C}_{\pi^\gamma} \frac{\operatorname{rank}(M)(m_1+m_2)[\log n + \ell^2(M)]}{\lambda} \right] + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda}$$

where the constant $\mathcal{C}_{\pi^\gamma}$ is explicitly known, and depends only on the form of prior π^γ (Gamma, or Inverse-Gamma) and of its hyperparameters.

An exact value for $\mathcal{C}_{\pi^\gamma}$ can be deduced from the proof. It is thus clear that the algorithm performs a trade-off between the fit to the data, through the term $r_n^h(M)$, and the rank of M.

In addition to empirical bounds, it is necessary to provide so-called theoretical bounds, that will prove that the risk of $\widetilde{\rho}_\lambda$ will indeed converge to the Bayes risk when the sample size grows. It is the goal of the next subsection.

3.3.2 Theoretical Bound

For this type of theoretical analysis, it is common in classification to make an additional assumption on $\mathbf{P}$ which leads to an easier task and therefore to better rates of convergence. We propose a definition adapted from [Mammen and Tsybakov, 1999].

Definition 3.2. *Mammen and Tsybakov margin assumption is satisfied when there is a constant C such that, for any matrix M:*

$$\mathbb{E}\left[\left(\mathbb{1}_{YM_X\leq 0} - \mathbb{1}_{YM_X^B\leq 0}\right)^2\right] \leq C[\mathbf{R}(M) - \overline{\mathbf{R}}].$$

It is known that if there is a constant $t > 0$ such that $\mathbb{P}(0 < |\eta(X)| < t) = 0$ then the margin assumption is satisfied with some C that depends on t. For example, in the noiseless case where $Y = M_X^B$ almost surely, which corresponds to $t = 1$, then

$$\begin{aligned}\mathbb{E}\left[\left(\mathbb{1}_{YM_X\leq 0} - \mathbb{1}_{YM_X^B\leq 0}\right)^2\right] &= \mathbb{E}\left[\mathbb{1}^2_{YM_X\leq 0}\right] = \mathbb{E}\left[\mathbb{1}_{YM_X\leq 0}\right] \\ &= \mathbf{R}(M) = \mathbf{R}(M) - \overline{\mathbf{R}},\end{aligned}$$

so the margin assumption is satisfied with $C = 1$.

We are now ready to state our theoretical bound. It makes a link between the integrated risk of the estimator and the lowest possible risk, which is reached by the Bayes classifier M^B. In opposition to the empirical bound, it involves non-observable quantities, depending on M^B, in the right-hand side.

Theorem 3.2. *Assume that Mammen and Tsybakov assumption is satisfied for a given constant $C > 0$. Then, for any $\epsilon \in (0, 1)$ and for $\lambda = sn/C$, $s \in (0, 1)$, with probability at least $1 - 2\epsilon$,*

$$\begin{aligned}\int \mathbf{R} d\widetilde{\rho}_\lambda \leq\ & 2(1 + 3s)\overline{\mathbf{R}} \\ & + \mathcal{C}_{C,s,\pi^\gamma}\left(\frac{\operatorname{rank}(M^B)(m_1 + m_2)[\log n + \ell^2(M^B)] + \log\left(\frac{1}{\epsilon}\right)}{n}\right)\end{aligned}$$

where $\mathcal{C}_{C,s,\pi^\gamma}$ is known and depends only on the s, C and π^γ.

Note the adaptive nature of this result, in the sense that the estimator does *not* depend on $\operatorname{rank}(M^B)$. Clearly, when $\operatorname{rank}(M^B)$ is small, the prediction error will be close to the Bayes error $\overline{\mathbf{R}}$ even for small sample size. This type of inequalitiy is often referred to as an 'oracle inequality' in the sense that our estimator behaves as well as if we knew the rank of M^B through an oracle.

Corollary 3.2. *In the noiseless case $Y = \text{sign}(M_X^B)$ a.s., for any $\epsilon > 0$ and for $\lambda = 2n$, with probability at least $1 - 2\epsilon$,*

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \mathcal{C}'_{\pi^\gamma} \left[\frac{\text{rank}(M^B)(m_1 + m_2)[\log n + \ell^2(M^B)] + \log \frac{1}{\epsilon}}{n} \right] \quad (3.7)$$

where $\mathcal{C}'_{\pi^\gamma} = \mathcal{C}_{1, \frac{1}{4}, 1, \pi^\gamma}$.

Remark 1. *Note that an empirical inequality comparable to Corollary 3.1 appears in [Srebro et al., 2005]. In both cases, the dependance of the bounds with respect to n is $1/\sqrt{n}$ (take $\lambda = \sqrt{n}$ in Corollary (3.1)). One notable difference is that our bound also provides an explicit dependance to the rank, which is not the case in [Srebro et al., 2005].*

In addition to this, theoretical inequalities like Theorem (3.2) and Corollary (3.2) are completely new results. They allow to compare the out-of-sample error of our predictor to the optimal one. They show that the rate is $\text{rank}(M^B)(m_1 + m_2)/n$ up to log terms. This can not be improved as this rate is known to be minimax optimal [Alquier et al., 2017].

Remark 2. *Determining the tuning parameter λ is not an easy task in practice: even though there are values that lead to the theoretical bounds, it is more efficient in practice to use cross-validation. We used this technique in the empirical results section.*

3.4 Algorithm

3.4.1 General Algorithm

The minimization problem (3.6) that defines our VB approximation is not straightforward:

$$\min_{L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \ldots, \rho^{\gamma_k}} AVB(L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \ldots, \rho^{\gamma_k}, \lambda).$$

When v^L, v^R and all the ρ^{γ_k}'s are fixed, this is actually the canonical example of so-called biconvex problems: it is convex with respect to L^0, and with respect to R^0, but not with respect to the pair (L^0, R^0). Such problems are notoriously difficult. In this case, alternating blockwise optimization seems to be an acceptable strategy. While there is no guarantee that the algorithm will not get stuck in a local minimum (or even in a singular point that is actually not a minimum), it seems to give very

good results in practice, and no efficient alternative is available. We refer the reader to the discussion in Subsection 9.2 page 76 [Boyd et al., 2011] for more details on this problem.

Our strategy is as follows. We update iteratively $L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \ldots, \rho^{\gamma_k}$: for L^0 and R^0 we use a gradient step, while for $v^L, v^R, \rho^{\gamma_1}, \ldots, \rho^{\gamma_k}$ an explicit minimization is available. The details for the mean-field optimization (that is, w.r.t. $\rho^{\gamma_1}, \ldots, \rho^{\gamma_k}$) are given in Subsection 3.4.2. See Algorithm 3.1 for the general version of the algorithm.

Algorithm 3.1 Variational Approximation with Hinge Loss

Require: $\epsilon, (\eta_t)_{t\in\mathbb{N}}, L_0^0, R_0^0, v_0^L, v_0^R, \rho_0^\gamma$
  $t \leftarrow 0$
  repeat
    $t \leftarrow t+1$
    $L_t^0 \leftarrow L_{t-1}^0 - \eta_t \frac{\partial AVB}{\partial L^0}(L_{t-1}^0, R_{t-1}^0, v_{t-1}^L, v_{t-1}^R, \rho_{t-1}^{\gamma_1}, \ldots, \rho_{t-1}^{\gamma_K}, \lambda)$
    $R_t^0 \leftarrow R_{t-1}^0 - \eta_t \frac{\partial AVB}{\partial R^0}(L_t^0, R_{t-1}^0, v_{t-1}^L, v_{t-1}^R, \rho_{t-1}^{\gamma_1}, \ldots, \rho_{t-1}^{\gamma_K}, \lambda)$
    $v_t^L \leftarrow \arg\min_{v^L} AVB(L_t^0, R_t^0, v^L, v_{t-1}^R, \rho_{t-1}^{\gamma_1}, \ldots, \rho_{t-1}^{\gamma_K}, \lambda)$
    $v_t^R \leftarrow \arg\min_{v^R} AVB(L_t^0, R_t^0, v_t^L, v^R, \rho_{t-1}^{\gamma_1}, \ldots, \rho_{t-1}^{\gamma_K}, \lambda)$
    $(\rho_t^{\gamma_1}, \ldots, \rho_t^{\gamma_K}) \leftarrow \arg\min_{\rho^{\gamma_1},\ldots,\rho^{\gamma_K}} AVB(L_t^0, R_t^0, v_t^L, v_t^R, \rho^{\gamma_1}, \ldots, \rho^{\gamma_K}, \lambda)$
  until $\|L_t^0 (R_t^0)^\top - L_{t-1}^0 (R_{t-1}^0)^\top\|_F^2 \leq \epsilon$

3.4.2 Mean Field Optimization

As the pseudo-likelihood does not involve the parameters $(\gamma_1, \ldots, \gamma_K)$, the variational distribution can be optimized in the same way as in [Lim and Teh, 2007] where the noise is Gaussian. The general update formula is:

$$
\begin{aligned}
\rho^{\gamma_k}(\gamma_k) &\propto \exp \mathbb{E}_{\rho^{-\gamma_k}} \left(\log \pi(L, R, \gamma)\right) \\
&\propto \exp \mathbb{E}_{\rho^{-\gamma_k}} \left(\log[\pi(L|\gamma)\pi(R|\gamma)\pi^\gamma(\gamma)]\right) \\
&\propto \exp \left\{ \sum_{i=1}^{m_1} \mathbb{E}_{\rho^L} \left[\log \pi(L_{i,k}|\gamma_k)\right] + \sum_{j=1}^{m_2} \mathbb{E}_{\rho^R} \left[\log \pi(R_{j,k}|\gamma_k)\right] + \log \pi^\gamma(\gamma_k) \right\} \\
&\propto \exp \left\{ -\frac{m_1+m_2}{2} \log \gamma_k - \frac{1}{\gamma_k} \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right] + \log \pi^\gamma(\gamma_k) \right\}
\end{aligned}
$$

where $\rho^{-\gamma_k}$ stands for the marginal distribution of $(\gamma_{k'})_{k'\neq k}$ under ρ. The solution then depends on π^γ. In what follows we derive explicit formulas for ρ^{γ_k} according to the choice of π^γ: we remind the reader that π^γ could be either a Gamma distribution, or an Inverse-Gamma distribution.

Inverse-Gamma Prior

The conjugate prior for this part of the model is the inverse-Gamma distribution. The prior of γ_k is $\pi^\gamma = \Gamma^{-1}(\alpha, \beta)$ and its density is:

$$\pi^\gamma(\gamma_k; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \gamma_k^{-\alpha-1} \exp\left(-\frac{\beta}{\gamma_k}\right) \mathbb{1}_{\mathbb{R}^+}(\gamma_k).$$

The moments we need to develop the algorithm and to compute the empirical bound are:

$$\mathbb{E}_{\pi^\gamma}(\log \gamma_k) = \log \beta - \psi(\alpha), \text{ and } \mathbb{E}_{\pi^\gamma}(1/\gamma_k) = \frac{\alpha}{\beta},$$

where ψ is the digamma function. Therefore, we get:

$$\rho^{\gamma_k}(\gamma_k) \propto \exp\left\{ -\left(\frac{m_1+m_2}{2} + \alpha + 1\right) \log \gamma_k \right. \\ \left. - \frac{1}{\gamma_k}\left(\mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right] + \beta \right) \right\},$$

so we can conclude that:

$$\rho^{\gamma_k} = \Gamma^{-1}\left(\frac{m_1+m_2}{2} + \alpha, \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right] + \beta \right). \quad (3.8)$$

Gamma Prior

Even though it seems that this fact was not used in prior works on Bayesian matrix estimation, it is also possible to derive explicit formulas when the prior π^γ on γ_k's is a $\Gamma(\alpha, \beta)$ distribution. In this case, ρ^{γ_k} is given by

$$\rho^{\gamma_k}(\gamma_k) \propto \exp\left\{ \left(\alpha - \frac{m_1+m_2}{2} - 1\right) \log \gamma_k - \beta\gamma_k - \frac{1}{\gamma_k} \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right] \right\}.$$

We remind the reader that the Generalized Inverse Gaussian distribution is a three-parameter family of distributions over $\mathbb{R}^{+*}$, written $GIG(a, b, \eta)$. Its density is given by:

$$f(x; a, b, \eta) = \frac{(a/b)^{\eta/2}}{2K_\eta(\sqrt{ab})} x^{\eta-1} \exp\left(-\frac{1}{2}(ax + bx^{-1})\right),$$

where K_λ is the modified Bessel function of second kind.

The variational distribution ρ^{γ_k} is in consequence $GIG(a_k, b_k, \eta_k)$ with:

$$a_k = 2\beta, \quad b_k = \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right], \quad \eta_k = \alpha - \frac{m_1 + m_2}{2}.$$

The moment we need in order to compute the variational distribution of L, R is:

$$\mathbb{E}_{\rho^{\gamma_k}} \left(\frac{1}{\gamma_k} \right) = \frac{K_{\eta_k - 1}(\sqrt{a_k b_k})}{K_{\eta_k}(\sqrt{a_k b_k})} \sqrt{\frac{a_k}{b_k}}.$$

3.5 Logistic Model

As mentioned in [Zhang, 2004], the hinge loss is not the only possible convex relaxation of the zero-one loss. The logistic loss $\text{logit}(u) = \log[1 + \exp(-u)]$ can also be used (even though it might lead to a loss in the rate of convergence of the risk the Bayes risk [Zhang, 2004]). This leads to the definitions:

$$\mathbf{R}^\ell(M) = \mathbb{E}_{\mathbf{P}} \left[\text{logit}(Y M_X) \right],$$
$$r_n^\ell(M) = \frac{1}{n} \sum_{k=1}^{n} \text{logit}(Y_k M_{X_k}).$$

In this case, the pseudo-likelihood $\exp(-\lambda r_n^\ell(M))$, if $\lambda = n$, is exactly equal to the likelihood of the logistic model:

$$Y|X = \begin{cases} 1 & \text{with probability } \sigma(M_X) \\ -1 & \text{with probability } 1 - \sigma(M_X) \end{cases}$$

where σ is the link function $\sigma(x) = \frac{\exp(x)}{1+\exp(x)}$. The likelihood is written $\Lambda(L, R) = \prod_{l=1}^{n} \sigma(Y_l (L R^\top)_{X_l})$. The prior distribution is exactly the same as in the previous sections and the object of interest is the posterior distribution:

$$\widehat{\rho}_l(d\theta) = \frac{\Lambda(L, R)\pi(d\theta)}{\int \Lambda(L, R)\pi(d\theta)}.$$

In order to deal with large matrices, it remains interesting to develop a variational Bayes algorithm. However it is not as simple as in the

quadratic loss model, see [Lim and Teh, 2007] in which the authors develop a mean field approximation, because the logistic likelihood leads to intractable update formulas. A common way to deal with this model is to maximize another quantity which is very close to the one we are interested in. The principle, coming from [Jaakkola and Jordan, 2000], is well explained in [Bishop, 2006] and an extended example can be found in [Latouche et al., 2015].

We consider the mean field approximation so the approximation is sought among the distributions ρ that are factorized

$$\rho(d\theta) = \prod_{i=1}^{m_1} \rho^{L_i}(dL_{i,\cdot}) \prod_{j=1}^{m_2} \rho^{R_j}(dR_{j,\cdot}) \prod_{k=1}^{K} \rho^{\gamma_k}(d\gamma_k).$$

We have the following decomposition, for all distribution ρ:

$$\log \int \Lambda(L,R)\pi(d\theta) = \mathcal{L}(\rho) + \mathcal{K}(\rho, \widehat{\rho}_l)$$
$$\text{with } \mathcal{L}(\rho) = \int \log\left(\frac{\Lambda(L,R)\pi(\theta)}{\rho(\theta)}\right)\rho(d\theta).$$

Since the left-hand side (called log-evidence) is fixed, minimizing the Kullback divergence w.r.t. ρ is the same as maximizing $\mathcal{L}(\rho)$. Unfortunately, this quantity is intractable. But a lower bound, which corresponds to a Gaussian approximation, is much more easier to optimize. We introduce the additional parameter $\xi = (\xi_l)_{l\in[n]}$.

Proposition 3.2. *For all $\xi \in \mathbb{R}^n$ and for all ρ,*

$$\mathcal{L}(\rho) \geq \int \log \frac{H(\theta,\xi)\pi(\theta)}{\rho(\theta)}\rho(d\theta) := \mathcal{L}(\rho,\xi)$$
$$\textit{where } \log H(\theta,\xi) = \sum_{l\in[n]} \left\{ \log\sigma(\xi_l) + \frac{Y_l(LR^\top)_{X_l} - \xi_l}{2} - \tau(\xi_l)\left[(LR^\top)^2_{X_l} - \xi_l^2\right] \right\}$$
$$\textit{and } \tau(x) = 1/(2x)(\sigma(x) - 1/2).$$

Hence the estimator is (even though ξ is not the parameter of interest):

$$(\widetilde{\rho}, \widetilde{\xi}) = \arg\min_{\rho\in\mathcal{F}} \mathcal{L}(\rho,\xi). \tag{3.9}$$

Bayes Algorithm

The lower bound $\mathcal{L}(\rho, \xi)$ is maximized with respect to ρ by the mean field algorithm. A direct calculation shows that the optimal distribution of each site (written with a star subscript) is given by:

$$\forall i \in [m_1], \log \rho_\star^{L_{i,\cdot}}(L_{i,\cdot}) = \int \log\left[H(\theta,\xi)\pi(\theta)\right] \rho^R(dR)\rho^\gamma(d\gamma) \prod_{i' \neq i} \rho(dL_{i',\cdot}) + \text{const}$$

$$\forall j \in [m_2], \log \rho_\star^{R_{j,\cdot}}(R_{j,\cdot}) = \int \log\left[H(\theta,\xi)\pi(\theta)\right] \rho^L(dL)\rho^\gamma(d\gamma) \prod_{j' \neq j} \rho(dR_{j',\cdot}) + \text{const}$$

As $\log H(\theta, \xi)$ is a quadratic form in $(L_{i,\cdot})_{i\in[m_1]}$ and $(R_{j,\cdot})_{j\in[m_2]}$, the variational distribution of each parameter is Gaussian and a direct calculation gives:

$$\rho_\star^{L_{i,\cdot}} = \mathcal{N}\left(\mathcal{M}_i^L, \mathcal{V}_i^L\right), \quad \rho_\star^{R_{j,\cdot}} = \mathcal{N}\left(\mathcal{M}_j^R, \mathcal{V}_j^R\right) \quad \text{where}$$

$$\mathcal{M}_i^L = \left(\frac{1}{2}\sum_{l\in\Omega_{i,\cdot}} Y_l \mathbb{E}_\rho\left[R_{j_l,\cdot}\right]\right)\mathcal{V}_i^L, \quad \mathcal{V}_i^L = \left(2\sum_{l\in\Omega_{i,\cdot}} \tau(\xi_l)\mathbb{E}_\rho[R_{j_l,\cdot}^\top R_{j_l,\cdot}] + \mathbb{E}_\rho\left[\text{diag}\left(\frac{1}{\gamma}\right)\right]\right)^{-1}$$

$$\mathcal{M}_j^R = \left(\frac{1}{2}\sum_{l\in\Omega_{\cdot,j}} Y_l \mathbb{E}_q\left[L_{i_l,\cdot}\right]\right)\mathcal{V}_j^R, \quad \mathcal{V}_j^R = \left(2\sum_{l\in\Omega_{\cdot,j}} \tau(\xi_l)\mathbb{E}_\rho[L_{i_l,\cdot}^\top L_{i_l,\cdot}] + \mathbb{E}_\rho\left[\text{diag}\left(\frac{1}{\gamma}\right)\right]\right)^{-1}$$

The variational optimization for γ is exactly the same as in the Hinge Loss setting (with both possible prior distributions Γ and Γ^{-1}). The optimization of the variational parameters is given by:

$$\forall l \in [n], \quad \widehat{\xi_l} = \sqrt{\mathbb{E}_\rho\left[(LR^\top)^2_{X_l}\right]}$$

3.6 Empirical Results

In this section we compare our methods to the other 1-bit matrix completion techniques on simulated and real datasets. It is worth noting that the low rank decomposition does not involve the same matrix: in our model, it affects the Bayesian classifier matrix; in logistic model, it concerns the parameter matrix. The estimate from our algorithm is $\widehat{M} = \mathbb{E}_{\widetilde{\rho_\lambda}}(L)\mathbb{E}_{\widetilde{\rho_\lambda}}(R)^\top$ and we focus on the zero-one loss in prediction. We first test the performances on simulated matrices and then experiment them on a real data set. We compare the four following models: (a) hinge loss with variational approximation (referred as *HL*), (b)

Bayesian logistic model with variational approximation (referred as *Logis.*), (c) the frequentist logistic model from [Davenport et al., 2014] (referred as *freq. Logis.*) and (d) the frequentist least squares model from [Mazumder et al., 2010] (referred as *SI* for SoftImpute). The former two are tested with both Gamma and Inverse-Gamma prior distributions. The hyperparameters are all tuned by cross validation. The parameter of the frequentist methods is a regularization parameter that is also tuned by cross-validation.

The choice of K in our methods is more difficult. A large K leads to more parameters to be estimated. This considerably slows down our algorithms. In the end, some (very large) values of K are not feasible in practice. Still, what we observe is that the prior leads to an adaptive estimator, in accordance with the theoretical results: when K is taken too large (but still small enough in order to keep the computations feasible), the additional parameters are shrunk to zero. Having observed this fact, we keep $K = 10$ in many simulations. Still, we added simulations with a larger value, $K = 50$, in order to show that this shrinkage effect indeed takes place.

From a theoretical perspective, the complexity of each step of Algorithm 3.1 is of order $(m_1 + m_2)K$. Each step only involves very simple calculations, no matrix operations. On the opposite, the methods that use the nuclear norm are very time-consuming because the complexity of the SVD is of order $m_1 m_2 \min(m_1, m_2)$. It is possible to use approximate SVD, but the method is more difficult to tune.

3.6.1 Simulated Data: Small Matrices

The goal is to assess the models with different scenarios of data generation. The general scheme of the simulations is as follows: the observations come from a 200×200 matrix and we pick randomly 20% of its entries. We set $K = 10$ in our algorithms. The observations are generated as:

$$Y_l = \operatorname{sign}\left(M_{i_l,j_l} + Z_l\right) B_l, \quad \text{where } M \in \mathbb{R}^{m \times m}, \quad (B_l, Z_l)_{l \in [n]} \text{ are iid.}$$

The noise term (B, Z) is such that $\mathbf{R}(M) = \overline{\mathbf{R}}$ and M has low rank noted r in the followings. The predictions are directly compared to M. Two types of matrices M are built: the *type A* corresponds to the favorable case to the hinge loss; the entries of M lie in $\{-1, +1\}$ [1]. The

1. The matrices are built by drawing r independent columns with only $\{-1, 1\}$. The remaining columns are randomly equal to one of the first r columns multiplied by a factor in $\{-1, 1\}$.

type B corresponds to the a more difficult classification problem because many entries of M are around 0: M is a product of two matrices with r columns where the entries are iid $\mathcal{N}(0,1)$. The noise term is specified in Table 3.1. Note that the example A3 may also be seen as a switch noise with probability $\frac{e}{1+e} \approx 0.73$. The experiments are done one time for each.

Table 3.1 – Type of Noise

Type	Name	B	Z	Y
1	No noise	$B = 1$ a.s.	$Z = 0$ a.s.	$Y_l = \text{sign}(M_{i_l,j_l})$ a.s
2	Switch	$B \sim 0.9\delta_1 + 0.1\delta_{-1}$	$Z = 0$ a.s.	$Y_l = \text{sign}(M_{i_l,j_l})$ w.p. 0.9
3	Logistic	$B = 1$ a.s.	$Z \sim$ Logistic	$Y_l = 1$ w.p. $\sigma(M_{i_l,j_l})$

Table 3.2 – Prediction Error on Simulated Observations - rank 3

Type	Logis.-G	Logis.-IG	HL-G	HL-IG	freq. Logis.	SI
A1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0 %
A2	0.5%	0.9%	0.1%	0.0%	0.5%	0.4%
A3	16.0%	15.9%	8.5%	8.5%	17.3%	17.3 %
B1	4.1%	4.0%	5.3%	5.8%	5.1%	5.6 %
B2	10.1%	10.1%	10.8%	10.6%	10.7%	10.8 %
B3	16.0%	16.0%	22.1%	21.3%	19.8%	19.8 %

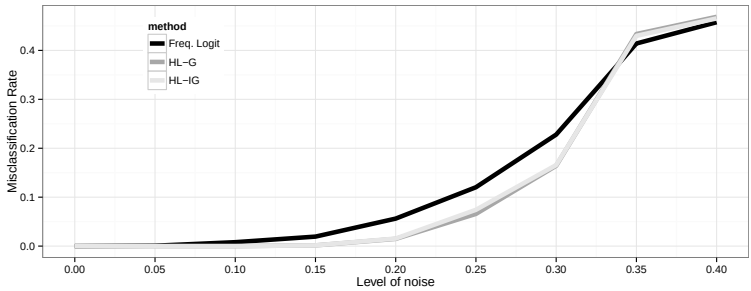
For rank 3 (see Table 3.2) and rank 5 matrices (see Table 3.3), the results of the Bayesian algorithms are very similar for both prior distributions and there is no evidence to favor a particular one. The results are better for the hinge loss method on type A observations and the difference of performance between models is very large for A3. On the opposite, the performance of the logistic model is better when the observations are generated from this model and when the parameter matrix is not separable. In comparison with the results from the frequentist approach, the variational approximation seems very good even though we have not at all any theoretical properties. For rank 5 matrices, the performances are worse but the meanings are the same as the rank 3 experiment.

The last simulation is a focus on the influence of the level of switch noise. On $A2$ type on rank 3 (Table 3.2), we see that 10% of corrupted entries is not enough to not almost perfectly recover the Bayesian classifier matrix. We challenge the frequentist program as well. The results are clear, see Figure 3.1: the hinge loss method is better almost everywhere.

Table 3.3 – Prediction Error on Simulated Matrices - rank 5

Type	Logis.-G	Logis.-IG	HL-G	HL-IG	freq. Logis.	SI
A1	0.01%	0.01%	0.0%	0.0%	0.01%	0.0 %
A2	4.4%	3.1%	0.54%	0.55%	3.1%	2.8 %
A3	32.5%	33.1%	27,0%	26.7%	30.1%	30.0 %
B1	7.8%	7.8%	9.4%	10.4%	9.0%	9.6 %
B2	17.3%	17.3%	17.9%	18.1%	18.3%	18.4 %
B3	21.5%	21.4%	24.4%	22.9%	22.1%	22.2 %

Figure 3.1 – Results on Simulated A2 Matrices with different levels of noise - rank 3



For a noise up to 25%, which means that one fourth of observed entries are corrupted, it is possible to get a very good predictor with less than 10% of misclassification error. It is getting worse when the level of noise increases and the problem becomes almost impossible for noise greater than 30%.

3.6.2 Simulated Data: Large Matrices

The second experiment involves larger matrices in order to assess the efficiency of the Bayesian methods on large dataset. The observations come now from a $2,000 \times 2,000$ matrix, 10% are observed randomly. The base matrix has now rank 10. It is worth noting that the matrix to be recovered is 100 times larger than in the first example. For the Bayesian methods, we fix $K = 50$. On the other hand, the frequentist methods need a singular value decomposition, which is very time consuming for such a large matrix. Exactly the same observation generation procedure

(Table 3.1) is used and six experiments are done (Table 3.4).

The results are in the line with the previous section: all the methods give very similar results except for high level of switch noise (matrix A3). The gap between the proposed methods and the logistic model is quite large and the algorithms using hinge loss perform almost perfectly. On the other hand, the hinge loss models perform well on logistic data and it is worth noting that the Bayesian logistic models perform better for logistic model with low level of noise (B1).

Table 3.4 – Prediction Error on Simulated Matrices - rank 10

Type	Logis.-G	Logis.-IG	HL-G	HL-IG	freq. Logis.	SI
A1	0 %	0 %	0 %	0 %	0 %	0 %
A2	0.1 %	0.3 %	0 %	0 %	0.1%	0.1 %
A3	9 .6 %	9.9 %	1.8 %	1.8 %	9.9%	9.9 %
B1	3.7 %	4.1 %	8.2 %	6.6 %	6.2%	7.4 %
B2	11 %	11 %	11.3%	10.6%	11.8%	12.0 %
B3	10.4 %	10.2 %	11.6%	11.3%	11.0%	11.2 %

3.6.3 Real Data set: MovieLens

The last experiment, presented in Table 3.5, involves the well known MovieLens [2] data set. It has already been used by [Davenport et al., 2014] and we follow them for the study. The ratings lie between 1 to 5 so we split them into binary data between good ratings (above the mean which is 3.5) and bad ones. The low rank assumption is usual in this case because it is expected that the taste of a particular user is related to only few hidden parameters. The smallest data set contains 100,000 ratings from 943 users and 1682 movies so we use 95,000 of them as a training set and the 5,000 remaining as the test set. The performances are very similar between the frequentist logistic model from [Davenport et al., 2014] and the hinge loss model. The performances seem slightly worse for the Bayesian logistic model but it is hard to favor a particular model at this stage (note that the difference between 0.28 and 0.27 is not significant on 5000 observations).

2. Available at http://grouplens.org/datasets/movielens/100k/

Table 3.5 – Misclassification rate on MovieLens 100k data set

Algorithm	HL-IG	HL-G	Logis.-G	Logis.-IG	Freq. Logis.
misclassif. rate	0.28	0.29	0.32	0.32	0.27

3.7 Discussion

We undertake the 1-bit matrix completion problem with classification tools and we are able to derive PAC-bounds on the risk and an efficient algorithm to compute the estimator. The previous works only focused on GLM models, which is not the right way to establish distribution free risk bounds. This work relies on PAC-Bayesian framework and the pseudo-posterior distribution is approximated by a variational algorithm. In practice, it is able to deal with large matrices. We also derive a variational approximation of the posterior distribution in the Bayesian logistic model and it works very well in our examples.

The variational approximations look very promising in order to build algorithm which are able to deal with large data and this framework may be extended to more general models and other Machine Learning tools.

Acknowledgment

We would like to thank Vincent Cottet's PhD supervisor Professor Nicolas Chopin, for his kind support during the project and the three anonymous referees for their helpful and constructive comments.

3.8 Proofs

3.8.1 Proofs of Proposition 3.1 from Section 3.2

Proof of Proposition 3.1. Let $\rho = \rho_{L^0,R^0,v^L,v^R,\rho^{\gamma_1},\ldots,\rho^{\gamma_k}}$ be a distribution from $\mathcal{F}$. The first term in (3.4) is upper bounded by the Lipschitz property of the hinge loss:

$$\int r_n^h(LR^\top)\rho(dL,dR,d\gamma) = \frac{1}{n}\sum_{h=1}^{n}\int \left(1 - Y_h L_{i_h,\cdot}.R_{j_h,\cdot}^\top\right)_+ \rho(dL,dR,d\gamma)$$

$$\leq \frac{1}{n}\sum_{h=1}^{n}\left(\left(1 - Y_h L_{i_h,\cdot}^0.R_{j_h,\cdot}^{0\top}\right)_+ + \int |L_{i_h,\cdot}.R_{j_h,\cdot}^\top - L_{i_h,\cdot}^0.R_{j_h,\cdot}^{0\top}|\rho(dL,dR)\right)$$

$$\leq r_n^h\left(L^0R^{0\top}\right)+\frac{1}{n}\sum_{h=1}^{n}\sum_{k=1}^{K}\left[\sqrt{v_{i_h,k}^L\frac{2}{\pi}}\sqrt{v_{j_h,k}^R\frac{2}{\pi}}+|R_{j_h,k}^0|\sqrt{v_{i_h,k}^L\frac{2}{\pi}}+|L_{i_h,k}^0|\sqrt{v_{j_h,k}^R\frac{2}{\pi}}\right].$$

The second part (KL-divergence) can be explicitly calculated. Let $\rho^{L_{i,k}}$ denote the marginal distribution of $L_{i,k}$ under ρ. We define in the same way $\rho^{R_{j,k}}$. Also, $\pi\rho^{L_{i,k}|\gamma_k}$ denote the distribution of $L_{i,k}$ given γ_k under π, and we define in the same way $\pi\rho^{R_{j,k}|\gamma_k}$. Then we have

$$\begin{aligned}
\mathcal{K}(\rho,\pi)&=\sum_{k=1}^{K}\left[\sum_{i=1}^{m_1}\mathbb{E}_{\rho^{\gamma_k}}\left[\mathcal{K}(\rho^{L_{i,k}},\pi^{L_{i,k}|\gamma_k})\right]+\sum_{j=1}^{m_2}\mathbb{E}_{\rho^{\gamma_k}}\left[\mathcal{K}(\rho^{R_{j,k}},\pi^{R_{j,k}|\gamma_k})\right]+\mathcal{K}(\rho^{\gamma_k},\pi^{\gamma_k})\right]\\
&=\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_{\rho^{\gamma_k}}\left[\frac{1}{\gamma_k}\right]\left(\sum_{i=1}^{m_1}(v_{i,k}^L+(L_{i,k}^0)^2)+\sum_{j=1}^{m_2}(v_{j,k}^R+(R_{j,k}^0)^2)\right)\\
&-\frac{1}{2}\sum_{k=1}^{K}\left(\sum_{i=1}^{m_1}\log v_{i,k}^L+\sum_{j=1}^{m_2}\log v_{j,k}^R\right)+\sum_{k=1}^{K}\left[\mathcal{K}(\rho^{\gamma_k},\pi^{\gamma_k})+\frac{m_1+m_2}{2}\left(\mathbb{E}_{\rho^{\gamma_k}}\left[\log\gamma_k\right]-1\right)\right]\\
&=\frac{1}{2}\sum_{k=1}^{K}\left[\sum_{i=1}^{m_1}\left(\mathbb{E}_{\rho^{\gamma_k}}\left[\frac{1}{\gamma_k}\right](v_{i,k}^L+(L_{i,k}^0)^2)+\mathbb{E}_{\rho^{\gamma_k}}\left[\log\gamma_k\right]-\log v_{i,k}^L-1\right)\right.\\
&\left.+\sum_{j=1}^{m_2}\left(\mathbb{E}_{\rho^{\gamma_k}}\left[\frac{1}{\gamma_k}\right](v_{j,k}^R+(R_{j,k}^0)^2)+\mathbb{E}_{\rho^{\gamma_k}}\left[\log\gamma_k\right]-\log v_{j,k}^R-1\right)+2\mathcal{K}(\rho^{\gamma_k},\pi^{\gamma_k})\right].
\end{aligned}$$

■

3.8.2 Proofs of the results in Subsection 3.3.1

Proof of Theorem 3.1. As the indicator function is uniformly bounded by 1, we can use Lemma 5.1 in [Alquier et al., 2016]:

$$\left.\begin{array}{l}\int\mathbb{E}\exp\{\lambda[\mathbf{R}(LR^\top)-r_n(LR^\top)]\}\mathrm{d}\pi(R,L,\gamma)\\ \int\mathbb{E}\exp\{\lambda[r_n(LR^\top)-\mathbf{R}(LR^\top)]\}\mathrm{d}\pi(R,L,\gamma)\end{array}\right\}\leq\exp\left[\frac{\lambda^2}{2n}\right].$$

So, the assumptions of Theorem 4.1 in [Alquier et al., 2016] are satisfied and we obtain that, for any $\epsilon\in(0,1)$, with probability at least $1-\epsilon$ on the drawing of the sample, for any ρ in $\mathcal{F}$:

$$\begin{aligned}
\int\mathbf{R}d\rho&\leq\int r_n d\rho+\frac{\mathcal{K}(\rho,\pi)}{\lambda}+\frac{\lambda}{2n}+\frac{\log\frac{1}{\epsilon}}{\lambda}\\
&\leq\int r_n^h d\rho+\frac{\mathcal{K}(\rho,\pi)}{\lambda}+\frac{\lambda}{2n}+\frac{\log\frac{1}{\epsilon}}{\lambda}\ (\text{as } r_n^h\geq r_n),
\end{aligned}$$

$$\leq r_n^h\left(L^0R^{0\top}\right) + \mathcal{R}(\rho,\lambda) + \frac{\lambda}{2n} + \frac{\log\frac{1}{\epsilon}}{\lambda} \text{ (thanks to Proposition 3.1).}$$

We end the proof by minimizing the right-hand-side w.r.t $\rho \in \mathcal{F}$. ■

In order to prove Corollary 3.1, we need a preliminary result. For any $m_1 \times m_2$ matrix M with rank $r \in [K]$, we can write $M = LR^T$ where L is $m_1 \times K$, R is $m_2 \times K$ and, up to a reordering of the columns, $L_{\cdot,r+1} = \cdots = L_{\cdot,K} = 0$ and $R_{\cdot,r+1} = \cdots = R_{\cdot,K} = 0$. We denote by $\mathcal{B}(M)$ the set of such pairs of matrices (L,R) and

$$\mathcal{F}(M) = \{\rho \in \mathcal{F} : (\mathbb{E}_\rho(L), \mathbb{E}_\rho(R)) \in \mathcal{B}(M)\}.$$

Lemma 3.2. *There is a constant $\mathcal{C}_{\pi^\gamma}$ that depends only on the choice of the prior π^γ such that for any $\lambda \leq n$,*

$$\inf_{\rho\in\mathcal{F}(M)} \mathcal{R}(\rho,\lambda) \leq \mathcal{C}_{\pi^\gamma}\frac{\operatorname{rank}(M)(m_1+m_2)(\ell(M)^2+\log n)}{\lambda},$$

$\mathcal{C}_{\pi^\gamma}$ is explicitly known and depends only on the choice of the prior π^γ (Gamma, or Inverse-Gamma) and of its parameters.

It is obvious that when we combine Lemma 3.2 with Theorem 3.1 we obtain Corollary 3.1. It remains to prove the lemma.

Proof. Let M be fixed and for short let r denote $\operatorname{rank}(M)$. We remind that, by definition:

$$\begin{aligned}
\mathcal{R}(\rho,\lambda) = \frac{1}{n}\sum_{h=1}^{n}\sum_{k=1}^{K}&\left[\sqrt{v^L_{i_h,k}\frac{2}{\pi}}\sqrt{v^R_{j_h,k}\frac{2}{\pi}} + |R^0_{j_h,k}|\sqrt{v^L_{i_h,k}\frac{2}{\pi}} + |L^0_{i_h,k}|\sqrt{v^R_{j_h,k}\frac{2}{\pi}}\right]\\
&+\frac{1}{\lambda}\left(\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_\rho\left[\frac{1}{\gamma_k}\right]\left(\sum_{i=1}^{m_1}(v^L_{i,k}+(L^0_{i,k})^2)+\sum_{j=1}^{m_2}(v^R_{j,k}+(R^0_{j,k})^2)\right)\right.\\
&-\frac{1}{2}\sum_{k=1}^{K}\left(\sum_{i=1}^{m_1}\log v^L_{i,k}+\sum_{j=1}^{m_2}\log v^R_{j,k}\right)\\
&\left.+\sum_{k=1}^{K}\left[\mathcal{K}(\rho^{\gamma_k},\pi^\gamma)+\frac{m_1+m_2}{2}\left(\mathbb{E}_\rho\left[\log\gamma_k\right]-1\right)\right]\right).
\end{aligned}$$

We will now upper bound the infimum for a special choice for $\rho = \rho_{L^0,R^0,v^L,v^R,\rho^{\gamma_1},\ldots,\rho^{\gamma_k}}$ with $(L^0,R^0)\in\mathcal{B}(M)$: for all pairs (i,k) and (j,k') $v^L_{i,k} = v^R_{j,k} = v^0$ when $k,k' \leq r$ $v^L_{i,k} = v^R_{j,k} = v^1$ otherwise. The choice for

v^0 and v^1 will be given below. For γ, we fix two distributions ρ_γ^0 and ρ_γ^1 and fix $\rho^{\gamma_k} = \rho_\gamma^0$ for $k \leq r$ and $\rho^{\gamma_k} = \rho_\gamma^1$ otherwise. Then:

$$\inf_{\rho \in \mathcal{F}(M)} \mathcal{R}(\rho, \lambda) \leq \frac{2}{\pi}((K-r)v^1 + rv_0) + 2r\ell(M)\sqrt{\frac{2v^0}{\pi}} + \frac{1}{\lambda}\left(r\mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1, \pi^\gamma)\right)$$
$$+ \frac{m_1+m_2}{2\lambda}\Bigg\{ r \underbrace{\left[\mathbb{E}_{\rho_\gamma^0}\left[\frac{1}{\gamma_k}\right](v^0 + \ell^2(M)) + \mathbb{E}_{\rho_\gamma^0}\log\gamma_k - \log v^0 - 1\right]}_{A_1}$$
$$+ (K-r)\underbrace{\left[\mathbb{E}_{\rho_\gamma^1}\left[\frac{1}{\gamma_k}\right]v_1 + \mathbb{E}_{\rho_\gamma^0}\log\gamma_k - \log v^1 - 1\right]}_{A_2}\Bigg\}.$$

By actually choosing $\rho_\gamma^0 = \pi^\gamma|_{[1,1+\delta]}$ for some $\delta > 0$ and $\rho_\gamma^1 = \pi^\gamma|_{[v^1, v_1+\delta]}$, we obtain

$$A_2 \leq 1 - 1 + \log\frac{v^1+\delta}{v^1} \leq \frac{\delta}{v^1};$$
$$A_1 \leq v^0 + l^2 + \delta - \log v^0.$$

At this stage, we can set the free parameters v_0, v_1 and δ in order to reach the desired rate. The choices are: $v_1 = \frac{1}{n}, v_0 = \frac{1}{n^2}, \delta = \frac{r}{Kn}$. We finally have to upper bound $r\mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1, \pi^\gamma)$. The upper bound actually depends on the choice for π^γ. We consider three cases: the Gamma prior with $\alpha \geq 1$, with $\alpha < 1$ and then the inverse-Gamma prior.

Let us deal with the $\Gamma^{-1}(\alpha, \beta)$ prior first:

$$\begin{aligned}
&r\mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1, \pi^\gamma) - K\log\frac{1}{\delta} - K\log\frac{\Gamma(\alpha)}{\beta^\alpha} \\
&\leq r[(\alpha+1)\log(1+\delta) + \beta] + (K-r)[(\alpha+1)\log(v_1+\delta) + \frac{\beta}{v_1}] \\
&\leq r[(\alpha+1)\delta + \beta] + (K-r)\left[(\alpha+1)(\log v_1 + \frac{\delta}{v_1}) + \frac{\beta}{v_1}\right] \\
&\leq r[(\alpha+1)\frac{r}{Kn} + \beta] + K\left[(\alpha+1)(-\log n + 1) + n\beta\right]
\end{aligned}$$

Let's turn to the $\Gamma(\alpha, \beta)$ distribution with $\alpha \geq 1$:

$$r\mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1, \pi^\gamma) - K\log\frac{1}{\delta} - K\log\frac{\Gamma(\alpha)}{\beta^\alpha}$$

$$\leq r[\beta(1+\delta)] + (K-r)[-(\alpha-1)\log v_1 + \beta(v_1+\delta)]$$
$$\leq r[\beta(1+\delta)] + (K-r)[(\alpha-1)\log\frac{1}{v_1} + \beta(v_1+\delta)]$$
$$\leq 2r\beta + K[(\alpha-1)\log n + \frac{2\beta}{n}]$$

The last case is the $\Gamma(\alpha,\beta)$ distribution with $0<\alpha<1$:

$$r\mathcal{K}(\rho_\gamma^0,\pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1,\pi^\gamma) - K\log\frac{1}{\delta} - K\log\frac{\Gamma(\alpha)}{\beta^\alpha}$$
$$\leq r[-(\alpha-1)\log(1+\delta) + \beta(1+\delta)] + (K-r)[-(\alpha-1)\log(v_1+\delta) + \beta(v_1+\delta)]$$
$$\leq r[(1-\alpha)\delta + \beta(1+\delta)] + (K-r)[(1-\alpha)(\log v_1 + \frac{\delta}{v_1}) + \beta(v_1+\delta)]$$
$$\leq 2r\beta + r(1-\alpha)\frac{r}{Kn} + K[(1-\alpha)(-\log n+1) + \frac{2\beta}{n}]$$

In any case, as $\lambda \leq n$, when α and β are constant, the leading term is in $\frac{r(m_1+m_2)(\ell^2(M)+\log n)}{\lambda}$ so we can upper bound the whole by $\mathcal{C}_{\pi^\gamma}\frac{r(m_1+m_2)(\ell^2(M)+\log n)}{\lambda}$ where $\mathcal{C}_{\pi^\gamma}$ depends on α and β (and takes a different form depending on the case: Gamma or inverse-Gamma). ■

Note actually that from the previous proof we can provide more explicit forms for the bound in the three cases. We did not include this in the core of the paper, but we prove the following lemmas for the sake of completeness.

Lemma 3.3. *When* $\pi^\gamma = \Gamma(\alpha,\beta)$,

$$\inf_{\rho\in\mathcal{F}(M)}\mathcal{R}(\rho,\lambda) \leq \frac{1}{n}\left[\frac{4}{\pi}K + \sqrt{\frac{8}{\pi}}r\ell(M)\right] + \frac{r(m_1+m_2)}{2\lambda}\left[3+\ell^2(M)+2\log n\right]$$
$$+ \frac{K}{\lambda}\left[\log\frac{Kn}{r} + \log\frac{\Gamma(\alpha)}{\beta^\alpha} + \frac{r}{K}\left[(\alpha+1)\frac{r}{Kn} + \beta\right] + [(\alpha+1)(-\log n+1) + n\beta]\right].$$

Lemma 3.4. *When* $\pi^\gamma = \Gamma^{-1}(\alpha,\beta)$ *with* $\alpha \geq 1$,

$$\inf_{\rho\in\mathcal{F}(M)}\mathcal{R}(\rho,\lambda) \leq \frac{1}{n}\left[\frac{4}{\pi}K + \sqrt{\frac{8}{\pi}}r\ell(M)\right] + \frac{r(m_1+m_2)}{2\lambda}\left[3+\ell^2(M)+2\log n\right]$$
$$+ \frac{K}{\lambda}\left[\log\frac{Kn}{r} + \log\frac{\Gamma(\alpha)}{\beta^\alpha} + \frac{2r\beta}{K} + \left[(\alpha-1)\log n + \frac{2\beta}{n}\right]\right].$$

Lemma 3.5. *When* $\pi^\gamma = \Gamma^{-1}(\alpha,\beta)$ *with* $0<\alpha<1$,

$$\inf_{\rho\in\mathcal{F}(M)}\mathcal{R}(\rho,\lambda) \leq \frac{1}{n}\left[\frac{4}{\pi}K + \sqrt{\frac{8}{\pi}}r\ell(M)\right] + \frac{r(m_1+m_2)}{2\lambda}\left[3+\ell^2(M)+2\log n\right]$$

$$+\frac{K}{\lambda}\left[\log\frac{Kn}{r}+\log\frac{\Gamma(\alpha)}{\beta^{\alpha}}+\frac{2r\beta}{K}+(1-\alpha)\frac{r^2}{K^2n}+\left[(1-\alpha)(-\log n+1)+\frac{2\beta}{n}\right]\right].$$

3.8.3 Proofs of the results in Subsection 3.3.2

We first start with preliminary lemmas.

Lemma 3.6. *For $\epsilon > 0$, with probability at least $1-\epsilon$ and for every $s \in (0,1)$,*

$$\overline{r_n} \leq (1+s)\overline{\mathbf{R}} + \frac{1}{ns}\log\frac{1}{\epsilon}$$

Proof. Let $s \in (0,1)$, then

$$\begin{aligned}
\mathbb{E}\left(\exp[sn\overline{r_n}]\right) &= \prod_{h=1}^{n}\mathbb{E}\left(\exp\left[s\mathbb{1}(Y_hM^B_{X_h}<0)\right]\right)\\
&= \prod_{h=1}^{n}\mathbb{E}\left(\exp\left[s\mathbb{1}(Y_hM^B_{X_h}<0)+0\left(1-\mathbb{1}(Y_hM^B_{X_h}<0)\right)\right]\right)\\
&\leq \prod_{h=1}^{n}\left(\left(1-\mathbb{E}\left[\mathbb{1}(Y_hM^B_{X_h}<0)\right]\right)+e^s\mathbb{E}\left[\mathbb{1}(Y_hM^B_{X_h}<0)\right]\right)\\
&\leq \prod_{h=1}^{n}\left((1-\overline{\mathbf{R}})+e^s\overline{\mathbf{R}}\right) = \prod_{h=1}^{n}\left(1+\overline{\mathbf{R}}(e^s-1)\right)\\
&\leq \prod_{h=1}^{n}\exp\left(\overline{\mathbf{R}}(e^s-1)\right) = \exp\left(n\overline{\mathbf{R}}(e^s-1)\right).
\end{aligned}$$

Therefore, for $\epsilon \in (0,1)$:

$$\mathbb{E}\left[\exp\left(sn\overline{r_n}-n\overline{\mathbf{R}}(e^s-1)-\log\frac{1}{\epsilon}\right)\right]\leq\epsilon.$$

We use the fact that $\mathbb{E}[\exp U] \geq \mathbb{P}(U>0)$ (Markov's inequality) for any U so, with probability at least $1-\epsilon$:

$$\overline{r_n} \leq \frac{e^s-1}{s}\overline{\mathbf{R}} + \frac{\log\frac{1}{\epsilon}}{sn}$$

On $[0,1]$, $e^x \leq 1+x+x^2$, thus ending the proof. ∎

Lemma 3.7. *Assume that Mammen and Tsybakov assumption is satisfied for a certain constant C. Assume that $\lambda < 2n/C$. Then, for $\epsilon > 0$, with probability at least $1 - \epsilon$:*

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \overline{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ \inf_{\rho \in \mathcal{F}} \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) \right] - \overline{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\} \tag{3.10}$$

Proof. Assume that the Mammen and Tsybakov assumption is satisfied for a certain constant C. The zero-one loss is bounded then, from Bernstein's inequality (Theorem 10 page 37 in [Boucheron et al., 2013]) we get:

$$\begin{aligned} &\int \mathbb{E} \exp\{\lambda[\mathbf{R}(LR^\top) - \overline{\mathbf{R}}] - \lambda[r_n(LR^\top) - \overline{r}_n]\} \mathrm{d}\pi(L, R, \gamma) \\ &\leq \int \exp[C\lambda^2/(2n)[\mathbf{R}(LR^\top) - \overline{\mathbf{R}}]] \mathrm{d}\pi(L, R, \gamma). \end{aligned}$$

Apply Fubini's theorem to the inequality:

$$\mathbb{E} \int \exp\{(\lambda - C\lambda^2/(2n))[\mathbf{R}(LR^\top) - \overline{\mathbf{R}}] - \lambda[r_n(LR^\top) - \overline{r}_n]\}) \pi(d\theta) \leq 1$$

(we remind that $\theta = (L, R, \gamma)$ for short).

$$\mathbb{E} \exp \left\{ \sup_\rho \int [\lambda[\mathbf{R}(LR^\top) - \overline{\mathbf{R}}] - \lambda[r_n(LR^\top) - \overline{r}_n] - C\lambda^2/(2n)[\mathbf{R}(LR^\top) - \overline{\mathbf{R}}]] \rho(d\theta) - \mathcal{K}(\rho, \pi) \right\} \leq 1.$$

Using Markov's inequality,

$$\mathbb{P} \left(\sup_\rho \int [(\lambda - C\lambda^2/(2n))[\mathbf{R}(LR^\top) - \overline{\mathbf{R}}] - \lambda[r_n(LR^\top) - \overline{r}_n]] \rho(d\theta) - \mathcal{K}(\rho, \pi) + \log \epsilon > 0 \right) \leq \epsilon.$$

Then take the complementary of this event and we get that with probability at least $1 - \epsilon$:

$$\forall \rho, \qquad (\lambda - C\lambda^2/(2n)) \int [\mathbf{R}(LR^\top) - \overline{\mathbf{R}}] \rho(d\theta) \leq \lambda \int [r_n(LR^\top) - \overline{r}_n] \rho(d\theta) + \mathcal{K}(\rho, \pi) + \log \frac{1}{\epsilon}$$

Now, note that

$$\begin{aligned} &(\lambda - C\lambda^2/(2n)) \left[\int \mathbf{R} d\rho - \overline{\mathbf{R}} \right] \leq \lambda \left[\int r_n d\rho - \overline{r}_n \right] + \mathcal{K}(\rho, \pi) + \log \left(\frac{1}{\epsilon} \right) \\ \Rightarrow &(\lambda - C\lambda^2/(2n)) \left[\int \mathbf{R} d\rho - \overline{\mathbf{R}} \right] \leq \lambda \left[\int r_n^h d\rho + \frac{1}{\lambda} \mathcal{K}(\rho, \pi) \right] - \lambda \overline{r}_n + \log \left(\frac{1}{\epsilon} \right) \\ \Rightarrow &(\lambda - C\lambda^2/(2n)) \left[\int \mathbf{R} d\rho - \overline{\mathbf{R}} \right] \leq \lambda \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) - \overline{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right] \end{aligned}$$

As it stands for all ρ then the right hand side can be minimized and the minimizer over $\mathcal{F}$ is $\widetilde{\rho}_\lambda$. Thus we get, when $\lambda < 2n/C$,

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \overline{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ \inf_{\rho \in \mathcal{F}} \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) \right] - \overline{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\}$$

■

We are now ready for the proofs.

Proof of Theorem 3.2. We apply Lemma 3.7 and, as we have $\mathcal{F}(M^B) \subset \mathcal{F}$,

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \overline{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ \inf_{\rho \in \mathcal{F}(M)} \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) \right] - \overline{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\} \tag{3.11}$$

As by definition, all the entries of M^B are in $\{-1, 1\}$, $r_n^h(M^B) = 2\overline{r_n}$ and then, by Lemma 3.2:

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \overline{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ 2\overline{r_n} + \mathcal{C}_{\pi^\gamma} \frac{\operatorname{rank}(M^B)(m_1 + m_2)(\ell^2(M) + \log n)}{\lambda} + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\}$$

Then, we use Lemma 3.7 to get, with probability at least $1 - 2\varepsilon$,

$$\int \mathbf{R} d\widetilde{\rho}_\lambda \leq \overline{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ 2(1+s)\overline{\mathbf{R}} + \frac{1}{ns} \log \frac{1}{\epsilon} + \mathcal{C}_{\pi^\gamma} \frac{\operatorname{rank}(M^B)(m_1 + m_2)(\ell^2(M) + \log n)}{\lambda} + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\}$$

To end up the proof, we have to take $\lambda = \frac{2cn}{C}$ with $c \in (0, 1/2)$. We thus have:

$$\frac{1}{1 - C\lambda/(2n)} = \frac{1}{1 - c} \leq 1 + 2c,$$

this ends the proof by taking $c = s/2$. ■

Proof of Corollary 3.2. As we are in the noiseless case, the margin assumption is satisfied with $C = 1$, and $\overline{\mathbf{R}} = 0$. ■

3.8.4 Detailed calculations for Subsection 3.5

Proof of Proposition 3.2. From [Jaakkola and Jordan, 2000], we have the following lower bound:

$$\forall (x,\xi) \in \mathbb{R}^2$$
$$\log\sigma(x) \geq \log\sigma(\xi) + \frac{x-\xi}{2} - \tau(\xi)(x^2-\xi^2) \quad \text{where} \quad \tau(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right)$$

The likelihood of one observation $y \in \{-1,1\}$ at point x is then lower bounded:

$$\forall \xi \in \mathbb{R}, \sigma(yx) \geq \sigma(\xi)\exp\left\{\frac{yx-\xi}{2} - \tau(\xi)(x^2-\xi^2)\right\} := h(yx,\xi).$$

Therefore, the likelihood of the model is lower bounded:

$$\begin{aligned}\forall \xi \in \mathbb{R}^n, \quad \Lambda(L,R) &= \prod_{l=1}^{n} \sigma(Y_l (LR^\top)_{X_l}) \\ &\geq \prod_{l=1}^{n} \sigma(\xi_l)\exp\left\{\frac{(LR^\top)_{X_l} - \xi_l}{2} - \tau(\xi_l)\left[(LR^\top)^2_{X_l} - \xi_l^2\right]\right\} := H(\theta,\xi).\end{aligned}$$

■

It is now easy to optimize $\mathcal{L}(\rho,\xi)$ with respect to ξ elementwise, which is the same as maximizing $H(\theta,\xi)$ elementwise and then each part $h(Y_l(LR^\top)_{X_l},\xi_l)$:

$$\begin{aligned}\widehat{\xi_l} &= \arg\max_{\xi_l} \int \log H(\theta,\xi)\rho(d\theta) = \arg\max_{\xi_l} \mathbb{E}_\rho\left[\log h(Y_l(LR^\top)_{X_l},\xi_l)\right] \\ &= \arg\max_{\xi_l}\left\{\log\sigma(\xi_l) - \frac{\xi_l}{2} - \tau(\xi_l)\left(\mathbb{E}_\rho\left[(LR^\top)^2_{X_l}\right] - \xi_{ij}^2\right)\right\}\end{aligned}$$

The maximum is reached at the zero of the derivative and we can conclude that:

$$\widehat{\xi_l}^2 = \mathbb{E}_\rho\left[(LR^\top)^2_{X_l}\right]$$

Chapter 4

Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions

With Pierre Alquier and Guillaume Lecué, submitted.

Abstract

We obtain estimation error rates and sharp oracle inequalities for regularization procedures of the form

$$\hat{f} \in \arg\min_{f \in F} \left(\frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i), Y_i) + \lambda \|f\| \right)$$

when $\|\cdot\|$ is any norm, F is a convex class of functions and ℓ is a Lipschitz loss function satisfying a Bernstein condition over F. We explore both the bounded and subgaussian stochastic frameworks for the distribution of the $f(X_i)$'s, with no assumption on the distribution of the Y_i's. The general results rely on two main objects: a complexity function, and a sparsity equation, that depend on the specific setting in hand (loss ℓ and norm $\|\cdot\|$).

As a proof of concept, we obtain minimax rates of convergence in the following problems: 1) matrix completion with any Lipschitz loss function, including the hinge and logistic loss for the so-called 1-bit matrix completion instance of the problem, and quantile losses for the general case, which enables to estimate any quantile on the entries of the matrix; 2) logistic LASSO and variants such as the logistic SLOPE; 3) kernel methods, where the loss is the hinge loss, and the regularization function is the RKHS norm.

4.1 Introduction

Many classification and prediction problems are solved in practice by regularized empirical risk minimizers (RERM). The risk is measured by a loss function and the quadratic loss function is the most popular function for regression. It has been extensively studied (cf. [Lecué and Mendelson, 2015b, Koltchinskii, 2011] among others). Still many other loss functions are popular among practitioners and are indeed extremely useful in specific situations.

First, let us mention the quantile loss in regression problems. The 0.5-quantile loss (also known as absolute or L_1 loss) is known to provide an indicator of conditional central tendency more robust to outliers than the quadratic loss. An alternative to the absolute loss for robustification is provided by the Huber loss. On the other hand, general quantile losses are used to estimate conditional quantile functions and are extremely useful to build confidence intervals and measures of risk, like *Values at Risk* (VaR) in finance.

Let us now turn to classification problems. The natural loss in this context, the so called 0/1 loss, leads very often to computationally intractable estimators. Thus, it is usually replaced by a convex loss function, such as the hinge loss or the logistic loss. A thorough study of convex loss functions in classification can be found in [Zhang, 2004].

All the aforementioned loss functions (quantile, Huber, hinge and logistic) share a common property: they are Lipschitz functions. This motivates a general study of RERM with any Lipschitz loss. Note that some examples were already studied in the literature: the $\|\cdot\|_1$-penalty with a quantile loss was studied in [Belloni and Chernozhukov, 2011] under the name "quantile LASSO" while the same penalty with the logistic loss was studied in [Van de Geer, 2008] under the name "logistic LASSO" (cf. [van de Geer, 2016]). The ERM strategy with Lipschitz proxys of

the 0/1 loss are studied in [Koltchinskii and Panchenko, 2002]. The loss functions we will consider in the examples of this paper are reminded below:

1. **hinge loss:** $\ell(y', y) = (1 - yy')_+ = \max(0, 1 - yy')$ for every $y \in \{-1, +1\}, y' \in \mathbb{R}$,
2. **logistic loss:** $\ell(y', y) = \log(1+\exp(-yy'))$ for every $y \in \{-1, +1\}$, $y' \in \mathbb{R}$;
3. **quantile regression loss:** for some parameter $\tau \in (0, 1)$, $\ell(y', y) = \rho_\tau(y - y')$ for every $y \in \mathbb{R}, y' \in \mathbb{R}$ where $\rho_\tau(z) = z(\tau - I(z \leq 0))$ for all $z \in \mathbb{R}$.

The two main theoretical results of the paper, stated in Section 4.2, are general in the sense that they do not rely on a specific loss function or a specific regularization norm. We develop two different settings that handle different assumptions on the design. In the first one, we assume that the family of predictors is subgaussian; in the second setting we assume that the predictors are uniformly bounded, this setting is well suited for classification tasks, including the 1-bit matrix completion problem. The rates of convergence rely on quantities that measure the complexity of the model and the size of the subdifferential of the norm.

To be more precise, the method works for any regularization function as long as it is a norm. If this norm has some sparsity inducing power, like the ℓ_1 or nuclear norms, thus the statistical bounds depend on the underlying sparsity around the oracle because the subdifferential is large. We refer these bounds as *sparsity dependent bounds*. If the norm does not induce sparsity, it is still possible to derive bounds that are now depending on the norm of the oracle because the subdifferential of the norm is very large in 0. We call it *norm dependent bounds* (aka "complexity dependent bounds" in [Lecué and Mendelson, 2015a]).

We study many applications that give new insights on diverse problems: the first one is a classification problem with logistic loss and LASSO or SLOPE regularizations. We prove that the rate of the SLOPE estimator is minimax in this framework. The second one is about matrix completion. We derive new excess risk bounds for the 1-bit matrix completion issue with both logistic and hinge loss. We also study the quantile loss for matrix completion and prove it reaches sharp bounds. We show several examples in order to assess the general methods as well as simulation studies. The last example involves the SVM and proves that "classic" regularization method with no special sparsity inducing power can be analyzed in the same way as sparsity inducing regularization methods.

A remarkable fact is that no assumption on the output Y is needed (while most results for the quadratic loss rely on an assumption of the tails of the distribution of Y). Neither do we assume any statistical model relating the "output variable" Y to the "input variable" X.

Mathematical background and notations. The observations are N i.i.d pairs $(X_i, Y_i)_{i=1}^N$ where $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ are distributed according to P. We consider the case where $\mathcal{Y}$ is a subset of $\mathbb{R}$ and let μ denote the marginal distribution of X_i. Let L_2 be the set of real valued functions f defined on $\mathcal{X}$ such that $\mathbb{E}f(X)^2 < +\infty$ where the distribution of X is μ. In this space, we define the L_2-norm as $\|f\|_{L_2} = (\mathbb{E}f(X)^2)^{1/2}$ and the L_∞ norm such that $\|f\|_{L_\infty} = \text{esssup}(|f(X)|)$. We consider a set of predictors $F \subseteq E$, where E is a subspace of L_2 and $\|\cdot\|$ is a norm over E (actually, in some situations we will simply have $F = E$, but in some natural examples we will consider bounded set of predictors, in the sense that $\sup_{f \in F} \|f\|_{L_\infty} < \infty$, which implies that F cannot be a subspace of L_2).

For every $f \in F$, the loss incurred when we predict $f(x)$, while the true output / label is actually y, is measured using a loss function ℓ: $\ell(f(x), y)$. For short, we will also use the notation $\ell_f(x, y) = \ell(f(x), y)$ the loss function associated with f. In this work, we focus on loss functions that are nonnegative, and Lipschitz, in the following sense.

Assumption 4.1 (Lipschitz loss function)**.** *For every $f_1, f_2 \in F$, $x \in \mathcal{X}$ and $y \in \mathbb{R}$, we have*

$$\left|\ell(f_1(x), y) - \ell(f_2(x), y)\right| \leq |f_1(x) - f_2(x)|.$$

Note that we chose a Lipschitz constant equal to one in Assumption 4.1. This can always be achieved by a proper normalization of the loss function. We define the oracle predictor as

$$f^* \in \arg\min_{f \in F} P\ell_f \text{ where } ^1 P\ell_f = \mathbb{E}\ell_f(X, Y)$$

1. Note that without any assumption on Y it might be that $P\ell_f = \mathbb{E}\ell_f(X, Y) = \infty$ for any $f \in F$. Our results remain valid in this case, but it is no longer possible to use the definition $f^* \in \arg\min_{f \in F} P\ell_f$. A general definition is as follows: fix any $f_0 \in F$. Note that for any $f \in F$, $\mathbb{E}[\ell_f(X, Y) - \ell_{f_0}(X, Y)]] \leq \mathbb{E}|(f - f_0)(X)| < \infty$ under the assumptions on F that will be stated in Section 4.2. It is then possible to define f^* as any minimizer of $\mathbb{E}[\ell_f(X, Y) - \ell_{f_0}(X, Y)]]$. This definition obviously coincides with the defintion $f^* \in \arg\min_{f \in F} P\ell_f$ when $P\ell_f$ is finite for some $f \in F$.

and (X, Y) is distributed like the (X_i, Y_i)'s. The objective of machine learning is to provide an estimator $\hat{f}$ that predicts almost as well as f^*. We usually formalize this notion by introducing the excess risk $\mathcal{E}(f)$ of $f \in F$ by

$$\mathcal{L}_f = \ell_f - \ell_{f^*} \text{ and } \mathcal{E}(f) = P\mathcal{L}_f.$$

Thus we consider the estimator of the form

$$\hat{f} \in \arg\min_{f \in F} \{P_N \ell_f + \lambda \|f\|\} \tag{4.1}$$

where $P_N \ell_f = (1/N) \sum_{i=1}^N \ell_f(X_i, Y_i)$ and λ is a regularization parameter to be chosen. Such estimators are usually called Regularized Empirical Risk Minimization procedure (RERM).

For the rest of the paper, we will use the following notations: let rB and rS denote the radius r ball and sphere for the norm $\|\cdot\|$, i.e. $rB = \{f \in E : \|f\| \leq r\}$ and $rS = \{f \in E : \|f\| = r\}$. For the L_2-norm, we write $rB_{L_2} = \{f \in L_2 : \|f\|_{L_2} \leq r\}$ and $rS_{L_2} = \{f \in L_2 : \|f\|_{L_2} = r\}$ and so on for the other norms.

Even though our results are valid in the general setting introduced above, we will develop the examples mainly in two directions that we will refer to *vector* and *matrix*. The *vector* case involves $\mathcal{X}$ as a subset of $\mathbb{R}^p$; we then consider the class of linear predictors, i.e. $E = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$. In this case, we denote for $q \in [1, +\infty]$, the l_q-norm in $\mathbb{R}^p$ as $\|\cdot\|_{l_q}$. The *matrix* case is also referred as the trace regression model: X is a random matrix in $\mathbb{R}^{m \times T}$ and we consider the class of linear predictors $E = \{\langle M, \cdot \rangle, M \in \mathbb{R}^{m \times T}\}$ where $\langle A, B \rangle = \text{Trace}(A^\top B)$ for any matrices A, B in $\mathbb{R}^{m \times T}$. The norms we consider are then, for $q \in [1, +\infty[$, the Schatten-q-norm for a matrix: $\forall M \in \mathbb{R}^{m \times T}, \|M\|_{S_q} = (\sum \sigma_i(M)^q)^{1/q}$ where $\sigma_1(M) \geq \sigma_2(M) \geq \cdots$ is the family of the singular values of M. The Schatten-1 norm is also called trace norm or nuclear norm. The Schatten-2 norm is also known as the Frobenius norm. The S_∞ norm, defined as $\|M\|_{S_\infty} = \sigma_1(M)$ is known as the operator norm.

The notation $\mathbf{C}$ will be used to denote positive constants, that might change from one instance to the other. For any real numbers a, b, we write $a \lesssim b$ when there exists a positive constant $\mathbf{C}$ such that $a \leq \mathbf{C}b$. When $a \lesssim b$ and $b \lesssim a$, we write $a \sim b$.

Proof of Concept. We now present briefly one of the outputs of our global approach: an oracle inequality for the 1-bit matrix completion

problem with hinge loss (we refer the reader to Section 4.4 for a detailed exposition of this example). While the general matrix completion problem has been extensively studied in the case of a quadratic loss, see [Koltchinskii et al., 2011, Lecué and Mendelson, 2015b] and the references therein, we believe that there is no satisfying solution to the so-called 1-bit matrix completion problem, that is for binary observations $\mathcal{Y} = \{-1, +1\}$. Indeed, the attempts in [Srebro et al., 2005, Cottet and Alquier, 2016] to use the hinge loss did not lead to rank dependent learning rates. On the other hand, [Lafond et al., 2014] studied RERM procedure using a statistical modeling approach and the logistic loss. While these authors prove optimal rates of convergence of their estimator with respect to the Frobenius norm, the excess classification risk, is not studied in their paper. However we believe that the essence of machine learning is to focus on this quantity – it is directly related to the average number of errors in prediction.

From now on we assume that $\mathcal{Y} = \{-1, +1\}$ and we consider the *matrix* framework. In matrix completion, we write the observed location as a mask matrix X: it is an element of the canonical basis $(E_{1,1}, \cdots, E_{m,T})$ of $\mathbb{R}^{m\times T}$ where for any $(p, q) \in \{1, \ldots, m\} \times \{1, \ldots, T\}$ the entry of $E_{p,q}$ is 0 everywhere except for the (p, q)-th entry where it equals to 1. We assume that there are constants $0 < \underline{c} \leq \bar{c} < \infty$ such that, for any (p, q), $\underline{c}/(mT) \leq \mathbb{P}(X = E_{p,q}) \leq \bar{c}/(mT)$ (this extends the uniform sampling distribution for which $\underline{c} = \bar{c} = 1$). These assumptions are encompassed in the following definition.

Assumption 4.2 (Matrix completion design)**.** *The sample size N is in $\{\min(m, T), \ldots, \max(m, T)^2\}$ and X takes value in the canonical basis $(E_{1,1}, \cdots, E_{m,T})$ of $\mathbb{R}^{m\times T}$. There are positive constants $\underline{c}, \bar{c}$ such that for any $(p, q) \in \{1, \ldots, m\} \times \{1, \ldots, T\}$,*

$$\underline{c}/(mT) \leq \mathbb{P}(X = E_{p,q}) \leq \bar{c}/(mT).$$

A predictor can be seen, for this problem, as the natural inner product with a real $m\times T$ matrix: $f(X) = \langle M, X\rangle = \mathrm{Tr}(X^\top M)$. The class F that we consider in Section 4.4 is the set of linear predictors where every entry of the matrix is bounded: $F = \{\langle \cdot, M\rangle : M \in bB_\infty\}$ where $bB_\infty = \{M = (M_{pq}) : \max_{p,q} |M_{pq}| \leq b\}$ for a specific b. This set is very common in matrix completion studies. But it is especially natural in this setting: indeed, the Bayes classifier, defined by $\overline{M} = \arg\min_{M\in\mathbb{R}^{m\times T}} \mathbb{E}\left(1 - Y\langle X, M\rangle\right)_+$, has entries in $[-1, 1]$. So, by taking $b = 1$ in the definition of F, we ensure that the oracle $M^* = \arg\min_{M\in\mathbb{R}^{m\times T}} \mathbb{E}\left(1 - Y\langle X, M\rangle\right)_+$ satisfies

$M^* = \bar{M}$, so there would be no point in taking $b > 1$. We will therefore consider the following RERM (using the hinge loss)

$$\widehat{M} \in \operatorname*{arg\,min}_{M \in B_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} (1 - Y_i \langle X_i, M \rangle)_+ + \lambda \|M\|_{S_1} \right) \tag{4.2}$$

where $\lambda > 0$ is some parameter to be chosen. We prove in Section 4.4 the following result.

Theorem 4.1. *Assume that Assumption 4.2 holds and there is $\tau > 0$ such that, for any $(p, q) \in \{1, \dots, m\} \times \{1, \dots, T\}$,*

$$\left| \overline{M}_{p,q} - \frac{1}{2} \right| \geq \tau. \tag{4.3}$$

There is a $c_0(\underline{c}, \bar{c}) > 0$, that depends only on $\underline{c}$ and $\bar{c}$, and that is formally introduced in Section 4.4 below, such that if one chooses the regularization parameter

$$\lambda = c_0(\underline{c}, \bar{c}) \sqrt{\frac{\log(m+T)}{N \min(m, T)}}$$

then, with probability at least

$$1 - \mathbf{C} \exp\left(-\mathbf{C} \operatorname{rank}(\overline{M}) \max(m, T) \log(m+T) \right), \tag{4.4}$$

the RERM estimator $\widehat{M}$ defined in (4.2) satisfies for every $1 \leq p \leq 2$,

$$\frac{1}{(mT)^{\frac{1}{p}}} \left\| \widehat{M} - \overline{M} \right\|_{S_p} \leq \mathbf{C} \operatorname{rank}(\overline{M})^{\frac{1}{p}} \sqrt{\frac{\log(m+T)}{N}} \frac{\max(m, T)^{1 - \frac{1}{p}}}{\min(m, T)^{\frac{1}{p} - \frac{1}{2}}}$$

and as a special case for $p = 2$,

$$\frac{1}{\sqrt{mT}} \left\| \widehat{M} - \overline{M} \right\|_{S_2} \leq \mathbf{C} \sqrt{\frac{\operatorname{rank}(\overline{M}) \max(m, T) \log(m+T)}{N}} \tag{4.5}$$

and its excess hinge risk is such that

$$\begin{aligned} \mathcal{E}_{hinge}(\widehat{M}) =& \mathbb{E}(1 - Y \left\langle X, \widehat{M} \right\rangle)_+ - \mathbb{E}(1 - Y \langle X, \overline{M} \rangle)_+ \\ &\leq \mathbf{C} \frac{\operatorname{rank}(\overline{M}) \max(m, T) \log(m+T)}{N} \end{aligned}$$

where the notation $\mathbf{C}$ is used for constants that might change from one instance to the other but depend only on $\underline{c}$, $\bar{c}$ and τ.

The excess hinge risk bound from Theorem 4.1 is of special interest as it can be related to the classic excess 0/1 risk. The excess 0/1 risk of a procedure is really the quantity we want to control since it measures the difference between the average number of mistakes of a procedure with the best possible theoretical classification rule. Indeed, let us define the 0/1 risk of M by $R_{0/1}(M) = \mathbb{P}[Y \neq \text{sign}(\langle M, X\rangle)]$. It is clear that $\overline{M} \in \arg\min_{M\in\mathbb{R}^{m\times T}} R_{0/1}(M)$. Then, it follows from Theorem 2.1 in [Zhang, 2004] that for some universal constant $c > 0$, for every $M \in \mathbb{R}^{m\times T}$,

$$R_{0/1}(M) - \inf_{M\in B_\infty} R_{0/1}(M) \leq c\mathcal{E}_{hinge}(M).$$

Therefore, the RERM from (4.2) for the choice of regularization parameter λ as in Theorem 4.1 satisfies with probability larger than in (4.4),

$$\mathcal{E}_{0/1}(\widehat{M}) = R_{0/1}(\widehat{M}) - \inf R_{0/1}(M) \leq \mathbf{C}\frac{\text{rank}(\overline{M})\max(m,T)\log(m+T)}{N} \tag{4.6}$$

where $\mathbf{C}$ depends on c, $\underline{c}$, $\bar{c}$ and τ. This yields a bound on the average of excess number of mistakes of $\widehat{M}$. To our knowledge such a prediction bound was not available in the literature on the 1-bit matrix completion problem. Let us compare Theorem 4.1 to the main result in [Lafond et al., 2014]. In [Lafond et al., 2014], the authors focus on the estimation error $\|\widehat{M} - M^*\|_{S_2}$, which seems less relevant for practical applications. In order to connect such a result to the excess classification risk, one can use the results in [Zhang, 2004] and in this case, the best bound that can be derived is of the order of $\sqrt{\text{rank}(M^*)\max(m,T)/N}$. Note that other authors focused on the classification error: Srebro et al. [2005] proved an excess error bound, but the bound does not depend on the rank of the oracle. The rate $\text{rank}(M^*)\max(m,T)/N$ derived from Theorem 4.1 for the 0/1-classification excess risk was only reached in [Cottet and Alquier, 2016], but in the very restrictive noiseless setting, which is equivalent to $\inf_M R_{0/1}(M) = 0$.

We hope that this example convinced the reader of the practical interest of the general study of $\hat{f}$ in (4.1). The rest of the paper is organized as follows. In Section 4.2 we introduce the concepts necessary to the general study of (4.1): namely, a complexity parameter, and a sparsity parameter. Thanks to these parameters, we define the assumptions necessary to our general results: the Bernstein condition, which is classic in learning theory to obtain fast rates [Lecué and Mendelson, 2015b], and a stochastic assumption on F (subgaussian, or bounded). The general

results themselves are eventually presented. The remaining sections are devoted to applications of our results to different estimation methods: the logistic LASSO and logistic SLOPE in Section 4.3, matrix completion in Section 4.4 and Support Vector Machines (SVM) in Section 4.5. For matrix completion, the optimality of the rates for the logistic and the hinge loss, that were not known, is also derived. In Section 4.6 we discuss the Bernstein condition for the three main loss functions of interest: hinge, logistic and quantile.

4.2 Theoretical Results

4.2.1 Applications of the main results: the strategy

The two main theorems in Sections 4.2.5 and 4.2.6 below are general in the sense that they allow the user to deal with any (nonnegative) Lipschitz loss function and any norm for regularization, but they involve quantities that depend on the loss and the norm. The aim of this Section is first to provide the definition of these objects and some hints on their interpretation, through examples. The theorems are then stated in both settings. Basically, the assumptions for the theorems are of three types:

1. the so-called Bernstein condition, which is a quantification of the identifiability condition. It basically tells how the excess risk $\mathcal{E}(f) = P\mathcal{L}_f = P(\ell_f - \ell_{f^*})$ is related to the L_2 norm $\|f - f^*\|_{L_2}$.
2. a stochastic assumption on the distribution of the $f(X)$'s for $f \in F$. In this work, we consider both a subgaussian assumption and a uniform boundedness assumption. Analysis of the two setups differ only on the way the "statistical complexity of F" is measured (cf. below the functions $r(\cdot)$ in Definition 4.8 and Definition 4.9).
3. finally, we introduce a sparsity parameter as in [Lecué and Mendelson, 2015b]. It reflects how the norm $\|\cdot\|$ used as a regularizer can induce sparsity - for example, think of the "sparsity inducing power" of the l_1-norm used to construct the LASSO estimator.

Given a scenario, that is a loss function ℓ, a random design X, a convex class F and a regularization norm, statistical results (exact oracle inequalities and estimation bounds w.r.t. the L_2 and regularization norms) for the associated regularized estimator together with the choice of the regularization parameter follow from the derivation of the three parameters (κ, r, ρ^*) as explained in the next box together with Theorem 4.2 and Theorem 4.3.

Application of the main results

1. find the **Bernstein parameter** $\kappa \geq 1$ and $A > 0$ associated to the loss and the class F;
2. compute the **Complexity function**

$$r(\rho) = \left[\frac{A\rho \mathrm{comp}(B)}{\sqrt{N}}\right]^{1/2\kappa}$$

   where $\mathrm{comp}(B)$ is defined either through the Gaussian mean width $w(B)$, in the subgaussian case, or the Rademacher complexity $\mathrm{Rad}(B)$, in the bounded case;
3. Compute the sub-differential $\partial\|\cdot\|(f^*)$ of $\|\cdot\|$ at the oracle f^* (or in the neighborhood $f^* + (\rho/20)B$ for approximately sparse oracles) and solve the **sparsity equation** "find ρ^* such that $\Delta(\rho^*) \geq 4\rho^*/5$".
4. Apply Theorem 4.2 in the subgaussian framework and Theorem 4.3 in the bounded framework. In each case, with large probability,

$$\left\|\hat{f} - f^*\right\| \leq \rho^*, \ \left\|\hat{f} - f^*\right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq \mathbf{C}\left[r(2\rho^*)\right]^{2\kappa}.$$

For the sake of simplicity, we present the two settings in different subsections with both the exact definition of the complexity function and the theorem. As the sparsity equation is the same in both settings, we define it before even though it involves the complexity function.

4.2.2 The Bernstein condition

The first assumption needed is called *Bernstein* assumption and is very classic in order to deal with Lipschitz loss.

Assumption 4.3 (Bernstein condition)**.** *There exists* $\kappa \geq 1$ *and* $A > 0$ *such that for every* $f \in F$, $\|f - f^*\|_{L_2}^{2\kappa} \leq AP\mathcal{L}_f$.

The most important parameter is κ and will be involved in the rate of convergence. As usual fast rates will be derived when $\kappa = 1$. In many situations, this assumption is satisfied and we present various cases in

Section 4.6. In particular, we prove that it is satisfied with $\kappa = 1$ for the logistic loss in both bounded and Gaussian framework, and we exhibit explicit conditions to ensure that Assumption 4.3 holds for the hinge and the quantile loss functions.

We call Assumption 4.3 a *Bernstein condition* following [Bartlett and Mendelson, 2006] and that it is different from the margin assumption from [Mammen and Tsybakov, 1999, Tsybakov, 2004]: in the so-called margin assumption, the oracle f^* in F is replaced by the minimizer $\overline{f}$ of the risk function $f \to P\ell_f$ over all measurable functions f, sometimes called the Bayes rules. We refer the reader to Section 4.6 and to the discussions in [Lecué and Mendelson, 2012] and Chapter 1.3 in [Lecué, 2011] for more details on the difference between the margin assumption and the Bernstein condition.

Remark 3. *The careful reader will actually realize that the proof of Theorem 4.2 and Theorem 4.3 requires only a weaker version of this assumption, that is: there exists $\kappa \geq 1$ and $A > 0$ such that for every $f \in \mathcal{C}$, $\|f - f^*\|_{L_2}^{2\kappa} \leq AP\mathcal{L}_f$, where $\mathcal{C}$ is defined in terms of the complexity function $r(\cdot)$ and the sparsity parameter ρ^* to be defined in the next subsections,*

$$\mathcal{C} := \left\{ f \in F : \|f - f^*\|_{L_2} \geq r(2\|f - f^*\|) \text{ and } \|f - f^*\| \geq \rho^* \right\}. \quad (4.7)$$

Note that the set $\mathcal{C}$ appears to play a central role in the analysis of regularization methods, cf. [Lecué and Mendelson, 2015b]. However, in all the examples presented in this paper, we prove that the Bernstein condition holds on the entire set F.

4.2.3 The complexity function $r(\cdot)$

The complexity function $r(\cdot)$ is defined by

$$\forall \rho > 0, \quad r(\rho) = \left[\frac{A\rho \mathrm{comp}(B)}{\sqrt{N}} \right]^{1/2\kappa}$$

where A is the constant in Assumption 4.3 and where $\mathrm{comp}(B)$ is a measure of the complexity of the unit ball B associated to the regularization norm. Note that this complexity measure will depend on the stochastic assumption of F. In the bounded setting, $\mathrm{comp}(B) = C\mathrm{Rad}(B)$ where C is an absolute constant and $\mathrm{Rad}(B)$ is the Rademacher complexity of B (whose definition will be reminded in Subsection 4.2.6). In the subgaussian setting, $\mathrm{comp}(B) = CLw(B)$ where C is an absolute constant, L is

the subgaussian parameter of the class $F - F$ and $w(B)$ is the Gaussian mean-width of B (here again, exact definitions of L and $w(B)$ will be reminded in Subsection 4.2.5).

Note that sharper (localized) versions of $r(\cdot)$ are provided in Section 4.8. However, as it is the simplest version that is used in most examples, we only introduce this version for now.

4.2.4 The sparsity parameter ρ^*

The size of the sub-differential of the regularization function $\|\cdot\|$ in a neighborhood of the oracle f^* will play as well a central role in our analysis. We recall now its definition: for every $f \in F$

$$\partial \|\cdot\| (f) = \{g \in E : \|f + h\| - \|f\| \geq \langle g, h \rangle \text{ for all } h \in E\} .$$

It is well-known that $\partial \|\cdot\| (f)$ is a subset of the unit sphere of the dual norm of $\|\cdot\|$ when $f \neq 0$. Note also that when $f = 0$, $\partial \|\cdot\| (f)$ is the entire unit dual ball, a fact we will also use in two situations, either when the regularization norm has no "sparsity inducing power" – in particular, when it is a smooth function as in the RKHS case treated in Section 4.5; or when one wants extra *norm dependent* upper bounds (cf. [Lecué and Mendelson, 2015a] for more details where these bounds are called *complexity dependent*) in addition to *sparsity dependent* upper bounds. In the latter, the statistical bounds that we get are the minimum between an error rate that depends on the notion of sparsity naturally associated to the regularization norm (when it exists) and an error rate that depends on $\|f^*\|$.

Definition 4.1 (From [Lecué and Mendelson, 2015b])**.** *The* ***sparsity parameter*** *is the function* $\Delta(\cdot)$ *defined by*

$$\Delta(\rho) = \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{g \in \Gamma_{f^*}(\rho)} \langle h, g \rangle$$

where $\Gamma_{f^*}(\rho) = \bigcup_{f \in f^* + (\rho/20)B} \partial \|\cdot\| (f)$.

Note that there is a slight difference with the definition of the *sparsity parameter* from [Lecué and Mendelson, 2015b] where there $\Delta(\rho)$ is defined taking the infimum over the sphere ρS intersected with a L_2-ball of radius $r(\rho)$ whereas in Definition 4.1, ρS is intersected with a L_2-ball of radius $r(2\rho)$. Up to absolute constants this has no effect on the behavior of $\Delta(\rho)$ and the difference comes from technical detains in our analysis (a peeling

argument that we use below whereas a direct homogeneity argument was enough in [Lecué and Mendelson, 2015b]).

In the following, estimation rates with respect to the regularization norm $\|\cdot\|$, the norm $\|\cdot\|_{L_2}$ as well as sharp oracle inequalities are given. All the convergence rates depend on a single radius ρ^* that satisfies the *sparsity equation* as introduced in [Lecué and Mendelson, 2015b].

Definition 4.2. *The radius ρ^* is any solution of the sparsity equation:*

$$\Delta(\rho^*) \geq (4/5)\rho^*. \tag{4.8}$$

Since ρ^* is central in the results and drives the convergence rates, finding a solution to the sparsity equation will play an important role in all the examples that we worked out in the following. Roughly speaking, if the regularization norm induces sparsity, a sparse element in $f^*+(\rho/20)B$ (that is an element f for which $\partial\|\cdot\|(f)$ is almost extremal – that is almost as large as the dual sphere) yields the existence of a small ρ^*. In this case, ρ^* satisfies the sparsity equation.

In addition, if one takes $\rho = 20\|f^*\|$ then $0 \in \Gamma_{f^*}(\rho)$ and since $\partial\|\cdot\|(0)$ is the entire dual ball associate to $\|\cdot\|$, one has directly that $\Delta(\rho) = \rho$ and so ρ satisfies the sparsity Equation (4.8). We will use this observation to obtain *norm dependent* upper bounds, i.e. rates of convergence depending on $\|f^*\|$ and that do not depend on any sparsity parameter. Such a bound holds for any norm; in particular, for norms with no sparsity inducing power as in Section 4.5.

4.2.5 Theorem in the subgaussian setting

First, we introduce the subgaussian framework (then we will turn to the bounded case in the next section).

Definition 4.3 (Subgaussian class)**.** *We say that a class of functions $\mathcal{F}$ is L-subgaussian (w.r.t. X) for some constant $L \geq 1$ when for all $f \in \mathcal{F}$ and all $\lambda \geq 1$,*

$$\mathbb{E}\exp\left(\lambda|f(X)|/\|f\|_{L_2}^2\right) \leq \exp\left(\lambda^2 L^2\right) \tag{4.9}$$

where $\|f\|_{L_2} = \left(\mathbb{E}f(X)^2\right)^{1/2}$.

We will use the following operations on sets: for any $F' \subset E$ and $f \in E$,

$$F' + f = \{f' + f : f' \in F'\}, \quad F' - F' = \{f_1' - f_2' : f_1', f_2' \in F'\}$$
$$\text{and } d_{L_2}(F') = \sup\left(\|f_1' - f_2'\|_{L_2} : f_1', f_2' \in F'\right).$$

Assumption 4.4. *The class $F - F$ is L-subgaussian.*

Note that there are many equivalent formulations of the subgaussian property of a random variable based on ψ_2-Orlicz norms, deviations inequalities, exponential moments, moments growth characterization, etc. (cf., for instance Theorem 1.1.5 in [Chafaï et al., 2012]). The one we should use later is as follows: there exists some absolute constant $\mathbf{C}$ such that $F - F$ is L-subgaussian if and only if for all $f, g \in F$ and $t \geq 1$,

$$\mathbb{P}[|f(X) - g(X)| \geq \mathbf{C}tL \|f - g\|_{L_2}] \leq 2\exp(-t^2). \tag{4.10}$$

There are several examples of subgaussian classes. For instance, when F is a class of linear functionals $F = \{\langle \cdot, t\rangle : t \in T\}$ for $T \subset \mathbb{R}^p$ and X is a random variable in $\mathbb{R}^p$ then $F - F$ is L-subgaussian in the following cases:

1. X is a Gaussian vector in $\mathbb{R}^p$,
2. $X = (x_j)_{j=1}^p$ has independent coordinates that are subgaussian, that is, there are constants $c_0 > 0$ and $c_1 > 0$ such that $\forall j$, $\forall t > c_0, \mathbb{P}[|x_j| \geq t(\mathbb{E}x_j^2)^{1/2}] \leq 2\exp(-c_1 t^2)$,
3. for $2 \leq q < \infty$, X is uniformly distributed over $p^{1/q}B_{l_q}$ (cf. [Barthe et al., 2005]),
4. $X = (x_j)_{j=1}^p$ is an unconditional vector (meaning that for every signs $(\epsilon_j)_j \in \{-1, +1\}^p$, $(\epsilon_j x_j)_{j=1}^p$ has the same distribution as $(x_j)_{j=1}^p$), $\mathbb{E}x_j^2 \geq c^2$ for some $c > 0$ and $\|X\|_{l_\infty} \leq R$ almost surely then one can choose $L \leq \mathbf{C}R/c$ (cf. [Lecué and Mendelson, 2013]).

In the *subgaussian framework*, a natural way to measure the *statistical complexity* of the problem is via Gaussian mean-width that we introduce now.

Definition 4.4. *Let H be a subset of L_2 and denote by d the natural metric in L_2. Let $(G_h)_{h\in H}$ be the canonical centered Gaussian process indexed by H (in particular, the covariance structure of $(G_h)_{h\in H}$ is given by d: $\left(\mathbb{E}(G_{h_1} - G_{h_2})^2\right)^{1/2} = \left(\mathbb{E}(h_1(X) - h_2(X))^2\right)^{1/2}$ for all $h_1, h_2 \in H$). The **Gaussian mean-width** of H (as a subset of L_2) is*

$$w(H) = \mathbb{E}\sup_{h\in H} G_h.$$

We refer the reader to Section 12 in [Dudley, 2002] for the construction of Gaussian processes in L_2. There are many natural situations

where Gaussian mean-widths can be computed. To familiarize with this quantity let us consider an example in the *matrix* framework. Let $H = \{\langle M, \cdot \rangle : \|M\|_{S_1} \leq 1\}$ be the class of linear functionals indexed by the unit ball of the S_1-norm and d be the distance associated with the Frobenius norm (i.e. $d(\langle \cdot, M_1 \rangle, \langle \cdot, M_2 \rangle) = d(M_1, M_2) = \|M_1 - M_2\|_{S_2}$) then

$$w(H) = w(B_{S_1}) = \mathbb{E} \sup_{\|M\|_{S_1} \leq 1} \langle \mathbb{G}, M \rangle = \mathbb{E} \|\mathbb{G}\|_{S_1}^* = \mathbb{E} \|\mathbb{G}\|_{S_\infty} \sim \sqrt{m+T}$$

where $\mathbb{G}$ is a standard Gaussian matrix in $\mathbb{R}^{m \times T}$, $\|\cdot\|_{S_1}^*$ is the dual norm of the nuclear norm which is the operator norm $\|\cdot\|_{S_\infty}$.

We are now in position to define the complexity parameter as announced previously.

Definition 4.5. *The **complexity parameter** is the non-decreasing function $r(\cdot)$ defined for every $\rho \geq 0$,*

$$r(\rho) = \left(\frac{ACLw(B)\rho}{\sqrt{N}} \right)^{\frac{1}{2\kappa}}$$

where κ, A are the Bernstein parameters from Assumption 4.3, L is the subgaussian parameter from Assumption 4.4 and $C > 0$ is an absolute constant (the exact value of C can be deduced from the proof of Proposition 4.6). The Gaussian mean-width $w(B)$ of B is computed with respect to the the metric associated with the covariance structure of X, i.e. $d(f_1, f_2) = \|f_1 - f_2\|_{L_2}$ for every $f_1, f_2 \in F$.

After the computation of the Bernstein parameter κ, the complexity function $r(\cdot)$ and the radius ρ^*, it is now possible to explicit our main result in the sub-Gaussian framework.

Theorem 4.2. *Assume that Assumption 4.1, Assumption 4.3 and Assumption 4.4 hold and let $C > 0$ from the definition of $r(\cdot)$ in Definition 4.5. Let the regularization parameter λ be*

$$\lambda = \frac{5}{8} \frac{CLw(B)}{\sqrt{N}}$$

and ρ^* satisfying (4.8)*. Then, with probability larger than*

$$1 - \mathbf{C} \exp\left(-\mathbf{C} N^{1/2\kappa} (\rho^* w(B))^{(2\kappa-1)/\kappa} \right) \tag{4.11}$$

we have

$$\left\|\hat{f}-f^*\right\| \leq \rho^*, \quad \left\|\hat{f}-f^*\right\|_{L_2} \leq r(2\rho^*) = \left[\frac{ACLw(B)2\rho^*}{\sqrt{N}}\right]^{1/2\kappa}$$
$$\text{and } \mathcal{E}(\hat{f}) \leq \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{CLw(B)2\rho^*}{\sqrt{N}}$$

where $\mathbf{C}$ *denotes positive constants that might change from one instance to the other and depend only on* A, κ, L *and* C.

Remark 4 (Deviation parameter)**.** *Replacing* $w(B)$ *by any upper bound does not affect the validity of the result. As a special case, it is possible to increase the confidence level of the bound by replacing* $w(B)$ *by* $w(B)+x$*: then, with probability at least*

$$1 - \mathbf{C}\exp\left(-\mathbf{C}N^{1/2\kappa}(\rho^*[w(B)+x])^{(2\kappa-1)/\kappa}\right)$$

we have in particular

$$\left\|\hat{f}-f^*\right\|_{L_2} \leq r(2\rho^*) = \left[\frac{ACL[w(B)+x]2\rho^*}{\sqrt{N}}\right]^{1/2\kappa}$$
$$\text{and } \mathcal{E}(\hat{f}) \leq \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{CL[w(B)+x]2\rho^*}{\sqrt{N}}.$$

Remark 5 (Norm and sparsity dependent error rates)**.** *Theorem 4.2 holds for any radius* ρ^* *satisfying the sparsity equation* (4.8)*. We have noticed in Section 4.2.4 that* $\rho^* = 20\|f^*\|$ *satisfies the sparsity equation since in that case* $0 \in \Gamma_{f^*}(\rho^*)$ *and so* $\Delta(\rho^*) = \rho^*$*. Therefore, one can apply Theorem 4.2 to both* $\rho^* = 20\|f^*\|$ *(this leads to* norm dependent *upper bounds) and to the smallest* ρ^* *satisfying the sparsity equation* (4.8) *(this leads to* sparsity dependent *upper bounds) at the same time. Both will lead to meaningful results (a typical example of such a combined result is Theorem 9.2 from [Koltchinskii, 2011] or Theorem 4.4 below).*

4.2.6 Theorem in the bounded setting

We now turn to the *bounded framework*; that is we assume that all the functions in F are uniformly bounded in L_∞. This assumption is very different in nature than the subgaussian assumption which is in fact a norm equivalence assumption (i.e. Definition 4.3 is equivalent to $\|f\|_{L_2} \leq \|f\|_{\psi_2} \leq L\|f\|_{L_2}$ for all $f \in \mathcal{F}$ where $\|\cdot\|_{\psi_2}$ is the ψ_2 Orlicz norm, cf. [Rao and Ren, 1991]).

Assumption 4.5 (Boundedness assumption)**.** *There exist a constant $b > 0$ such that for all $f \in F$, $\|f\|_{L_\infty} \leq b$.*

The main motivation to consider the *bounded setup* is for sampling over the canonical basis of a finite dimensional space like $\mathbb{R}^{m\times T}$ or $\mathbb{R}^p$. Note that this type of sampling is *stricto sensu* subgaussian, but with a constant L depending on the dimensions m and T, which yields suboptimal rates. This is the reason why the results in the bounded setting are more relevant in this situation. This is especially true for the 1-bit matrix completion problem as introduced in Section 4.1. For this example, the X_i's are chosen randomly in the canonical basis $(E_{1,1}, \cdots, E_{m,T})$ of $\mathbb{R}^{m\times T}$. Moreover, in that example, the class F is the class of all linear functionals indexed by bB_∞: $F = \{\langle \cdot, M\rangle : \max_{p,q} |M_{pq}| \leq b\}$ and therefore the study of this problem falls naturally in the bounded framework studied in this section.

Under the boundedness assumption, the natural way to measure the "statistical complexity" cannot be anymore characterized by Gaussian mean width. We therefore introduce another complexity parameter known as Rademacher complexities. This complexity measure has been extensively studied in the learning theory literature (cf., for instance, [Koltchinskii, 2006, 2011, Bartlett et al., 2005]).

Definition 4.6. *Let H be a subset of L_2. Let $(\epsilon_i)_{i=1}^N$ be N i.i.d. Rademacher variables (i.e. $\mathbb{P}[\epsilon_i = 1] = \mathbb{P}[\epsilon_i = -1] = 1/2$) independent of the X_i's. The **Rademacher complexity** of H is*

$$\mathrm{Rad}(H) = \mathbb{E} \sup_{f\in H} \Big| \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i f(X_i) \Big|.$$

Note that when $(f(X))_{f\in H}$ is a version of the isonormal process over L_2 (cf. Chapter 12 in [Dudley, 2002]) restricted to H then the Gaussian mean-width and the Rademacher complexity coincide: $w(H) = \mathrm{Rad}(H)$. But, in that case, H is not bounded in L_∞ and, in general, the two complexity measures are different.

There are many examples where Rademacher complexities have been computed (cf. [Mendelson, 2004]). Like in the previous *subgaussian* setting the statistical complexity is given by a function $r(\cdot)$ (we use the same name in the two *bounded* and *subgaussian* setups because this $r(\cdot)$ function plays exactly the same role in both scenarii even though it uses different notion of complexity).

Definition 4.7. *The **complexity parameter** is the non-decreasing function $r(\cdot)$ defined for every $\rho \geq 0$ by*

$$r(\rho) = \left(\frac{CA\mathrm{Rad}(B)\rho}{\sqrt{N}} \right)^{\frac{1}{2\kappa}}, \text{ where } C = \frac{1920}{7}.$$

Theorem 4.3. *Assume that Assumption 4.1, Assumption 4.3 and Assumption 4.5 hold. Let the regularization parameter λ be chosen as $\lambda = 720\mathrm{Rad}(B)/7\sqrt{N}$. Then, with probability larger than*

$$1 - \mathbf{C} \exp\left(-\mathbf{C} N^{1/2\kappa} (\rho^* \mathrm{Rad}(B))^{(2\kappa-1)/\kappa} \right) \tag{4.12}$$

we have

$$\left\| \hat{f} - f^* \right\| \leq \rho^*, \quad \left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\rho^*) = \left[\frac{CA\mathrm{Rad}(B)2\rho^*}{\sqrt{N}} \right]^{1/2\kappa}$$
$$\text{and } \mathcal{E}(\hat{f}) \leq \frac{r(2\rho^*)^{2\kappa}}{A} = \frac{C\mathrm{Rad}(B)2\rho^*}{\sqrt{N}},$$

where $\mathbf{C}$ denotes positive constants that might change from one instance to the other and depend only on A, b, κ and $r(\cdot)$ is the function introduced in Definition 4.7.

In the next Sections 4.3, 4.4 and 4.5 we compute $r(\rho)$ either in the subgaussian setup or in the bounded setup and solve the sparsity equation in various examples, showing the versatility of the main strategy.

4.3 Application to logistic LASSO and logistic SLOPE

The first example of application of the main results in Section 4.2 involves one very popular method developed during the last two decades in binary classification which is the Logistic LASSO procedure (cf. [Mak, 1999, Meier et al., 2008, Tian et al., 2008, Garcia-Magariños et al., 2010, Sabbe et al., 2013]).

We consider the *vector* framework, where $(X_1, Y_1), \ldots, (X_N, Y_N)$ are N i.i.d. pairs with values in $\mathbb{R}^p \times \{-1, 1\}$ distributed like (X, Y). Both bounded and subgaussian framework can be analyzed in this example. For the sake of shortness and since an example in the bounded case is

provided in the next section, only the subgaussian case is considered here and we leave the bounded case to the interested reader. We therefore shall apply Theorem 4.2 to get estimation and prediction bounds for the well known logistic LASSO and the new logistic SLOPE.

In this section, we consider the class of linear functional indexed by RB_{l_2} for some radius $R \geq 1$ and the logistic loss:

$$F = \{\langle \cdot, t\rangle : t \in RB_{l_2}\}, \ell_f(x,y) = \log(1+\exp(-yf(x))).$$

As usual the oracle is denoted by $f^* = \arg\min_{f\in F} \mathbb{E}\ell_f(X,Y)$, we also introduce t^* such that $f^* = \langle \cdot, t^*\rangle$.

4.3.1 Logistic LASSO

The logistic loss function is Lipschitz with constant 1, so Assumption 4.1 is satisfied. It follows from Proposition 4.2 in Section 4.6.1 that Assumption 4.3 is satisfied when the design X is the standard Gaussian variable in $\mathbb{R}^p$ and the considered class F. In that case, the Bernstein parameter is $\kappa = 1$, and we have $A = c_0/R^3$ for some absolute constant $c_0 > 0$ which can be deduced from the proof of Proposition 4.2. We consider the l_1 norm $\|\langle \cdot, t\rangle\| = \|t\|_{l_1}$ for regularization. We will therefore obtain statistical results for the RERM estimator $\widehat{f_L} = \langle \widehat{t_L}, \cdot\rangle$ that is defined by

$$\widehat{t_L} \in \underset{t\in RB_{l_2}}{\arg\min} \left(\frac{1}{N}\sum_{i=1}^{N} \log\left(1+\exp(-Y_i\langle X_i, t\rangle\right) + \lambda \|t\|_{l_1} \right)$$

where λ is a regularization parameter to be chosen according to Theorem 4.2.

The two final ingredients needed to apply Theorem 4.2 are 1) the computation of the Gaussian mean width of the unit ball B_{l_1} of the regularization function $\|\cdot\|_{l_1}$ 2) find a solution ρ^* to the sparsity equation (4.8).

Let us first deal with the complexity parameter of the problem. If one assumes that the design vector X is **isotropic**, i.e. $\mathbb{E}\langle X, t\rangle^2 = \|t\|_{l_2}^2$ for every $t \in \mathbb{R}^p$ then the metric naturally associated with X is the canonical l_2-distance in $\mathbb{R}^p$. In that case, it is straightforward to check that $w(B_{l_1}) \leq c_1\sqrt{\log p}$ for some (known) absolute constant $c_1 > 0$ and

so we define, for all $\rho \geq 0$,

$$r(\rho) = \mathbf{C}\left(\rho\sqrt{\frac{\log p}{N}}\right)^{1/2} \tag{4.13}$$

for the complexity parameter of the problem (from now and until the end of Section 4.3, the constants $\mathbf{C}$ depends only on L, C, c_0 and c_1).

Now let us turn to a solution ρ^* of the sparsity equation (4.8). First note that when the design is isotropic the sparsity parameter is the function

$$\Delta(\rho) = \inf\left\{\sup_{g\in\Gamma_{t^*}(\rho)} \langle h, g\rangle : h \in \rho S_{l_1} \cap r(2\rho)B_{l_2}\right\}$$

where $\Gamma_{t^*}(\rho) = t^* + (\rho/20)B_{l_1}$.

A first solution to the sparsity equation is $\rho^* = 20\,\|t^*\|_{l_1}$ because it leads to $0 \in \Gamma_{t^*}(\rho^*)$. This solution is called *norm dependent*.

Another radius ρ^* solution to the sparsity equation (4.8) is obtained when t^* is close to a sparse-vector, that is a vector with a small support. We denote by $\|v\|_0 := |\mathrm{supp}(v)|$ the size of the support of $v \in \mathbb{R}^p$. Now, we recall a result from [Lecué and Mendelson, 2015b].

Lemma 4.1 (Lemma 4.2 in [Lecué and Mendelson, 2015b])**.** *If there exists some $v \in t^* + (\rho/20)B_{l_1}$ such that $\|v\|_0 \leq c_0(\rho/r(\rho))^2$ then $\Delta(\rho) \geq 4\rho/5$ where c_0 is an absolute constant.*

In particular, we get that $\rho^* \sim s\sqrt{(\log p)/N}$ is a solution to the sparsity equation if there is a s-sparse vector which is $(\rho^*/20)$-close to t^* in l_1. This radius leads to the so-called *sparsity dependent* bounds.

After the derivation of the Bernstein parameter $\kappa = 1$, the complexity $w(B)$ and a solution ρ^* to the sparsity equation, we are now in a position to apply Theorem 4.2 to get statistical bounds for the Logistic LASSO.

Theorem 4.4. *Assume that X is a standard Gaussian vector in $\mathbb{R}^p$. Let $s \in \{1, \ldots, p\}$. Assume that there exists a s-sparse vector in $t^* + \mathbf{C}s\sqrt{(\log p)/N}B_{l_1}$. Then, with probability larger than $1-\mathbf{C}\exp(-\mathbf{C}s\log p)$, for every $1 \leq q \leq 2$, the logistic LASSO estimator $\widehat{t_L}$ with regularization parameter*

$$\lambda = \frac{5c_1CL}{8}\sqrt{\frac{\log p}{N}}$$

satisfies

$$\left\|\widehat{t_L} - t^*\right\|_{l_q} \leq \mathbf{C} \min\left(s^{1/q}\sqrt{\frac{\log p}{N}}, \|t^*\|_{l_1}^{1/q} \left(\frac{\log p}{N}\right)^{\frac{1}{2}-\frac{1}{2q}} \right)$$

and the excess logistic risk of $\widehat{t_L}$ *is such that*

$$\mathcal{E}_{logistic} = R(\widehat{t_L}) - R(t^*) \leq \mathbf{C} \min\left(\frac{s\log(p)}{N}, \|t^*\|_{l_1}\sqrt{\frac{\log(p)}{N}} \right).$$

Note that an estimation result for any l_q-norm for $1 \leq q \leq 2$ follows from results in l_1 and l_2 and the interpolation inequality $\|v\|_{l_q} \leq \|v\|_{l_1}^{-1+2/q} \|v\|_{l_2}^{2-2/q}$.

Estimation results for the logistic LASSO estimator in the generalized linear model have been obtained in [Van de Geer, 2008] under the assumption that the basis functions and the oracle are bounded. This assumption does not hold here since the *basis functions* – defined here by $\psi_k(\cdot) = \langle e_k, \cdot \rangle$ where $(e_k)_{k=1}^d$ is the canonical basis of $\mathbb{R}^p$ – are not bounded when the design is $X \sim \mathcal{N}(0, I_{d\times p})$. Moreover, we do not make the assumption that f^* is bounded in L_∞. Nevertheless, we recover the same estimation result for the l_2-loss and l_1-loss as in [Van de Geer, 2008]. But we also provide a prediction result since an excess risk bound is also given in Theorem 4.4.

Note that Theorem 4.4 recovers the classic rates of convergence for the logistic LASSO estimator that have been obtained in the literature so far. This rates is the minimax rate as long as $\log(p/s)$ behaves like $\log p$. This is indeed the case when $s \ll p$ which is the classic setup in high-dimensional statistics. But when s is proportional to p this rate is not minimax since there is a logarithmic loss. To overcome this issue we introduce a new estimator: the logistic SLOPE.

4.3.2 Logistic Slope

The construction of the logistic Slope is similar to the one of the logistic LASSO except that the regularization norm used in this case is the SLOPE norm (cf. [Su and Candès, 2016, Bogdan et al., 2015]): for every $t = (t_j) \in \mathbb{R}^p$,

$$\|t\|_{SLOPE} = \sum_{j=1}^{p} \sqrt{\log(ep/j)} t_j^\sharp \tag{4.14}$$

where $t_1^\sharp \geq t_2^\sharp \geq \cdots \geq 0$ is the non-increasing rearrangement of the absolute values of the coordinates of t and e is the base of the natural logarithm. Using this estimator with a regularization parameter $\lambda \sim 1/\sqrt{N}$ we recover the same result as for the Logistic LASSO case except that one can get, in that case, the optimal minimax rate for any $s \in \{1, \ldots, p\}$:

$$\sqrt{\frac{s}{N} \log\left(\frac{ep}{s}\right)}.$$

Indeed, it follows from Lemma 5.3 in [Lecué and Mendelson, 2015b] that the Gaussian mean width of the unit ball B_{SLOPE} associated with the SLOPE norm is of the order of a constant. The *sparsity dependent* radius satisfies

$$\rho^* \sim \frac{s}{\sqrt{N}} \log\left(\frac{ep}{s}\right) \tag{4.15}$$

as long as there is a s-sparse vector in $t^* + (\rho^*/20)B_{SLOPE}$. The *norm dependent* radius is as usual of order $\|t^*\|_{SLOPE}$. Then, the next result follows from Theorem 4.2. It improves the best known bounds on the logistic LASSO.

Theorem 4.5. *Assume that X is a standard Gaussian vector in $\mathbb{R}^p$. Let $s \in \{1, \ldots, p\}$. Assume that there exists a s-sparse vector in $t^* + (\rho^*/20)B_{SLOPE}$ for ρ^* as in (4.15). Then, with probability larger than $1 - \mathbf{C}\exp(-\mathbf{C}s\log(p/s))$, the logistic SLOPE estimator*

$$\widehat{t_S} \in \operatorname*{arg\,min}_{t \in RB_{l_2}} \left(\frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp(-Y_i \langle X_i, t\rangle\right) + \frac{\mathbf{C}}{\sqrt{N}} \|t\|_{SLOPE} \right)$$

satisfies

$$\left\|\widehat{t_S} - t^*\right\|_{SLOPE} \leq \mathbf{C} \min\left(\frac{s}{\sqrt{N}} \log\left(\frac{ep}{s}\right), \|t^*\|_{SLOPE} \right)$$

and

$$\left\|\widehat{t_S} - t^*\right\|_{l_2} \leq \mathbf{C} \min\left(\sqrt{\frac{s}{N} \log\left(\frac{ep}{s}\right)}, \sqrt{\frac{\|t^*\|_{SLOPE}}{\sqrt{N}}} \right)$$

and the excess logistic risk of $\widehat{t_S}$ is such that

$$\mathcal{E}_{logistic}(\widehat{t_S}) = R(\widehat{t_S}) - R(t^*) \leq \mathbf{C} \min\left(\frac{s \log ep/s}{N}, \|t^*\|_{l_1} \sqrt{\frac{\log ep/s}{N}} \right).$$

Let us comment on Theorem 4.5 together with the fact that we do not make any assumption on the output Y all along this work. Theorem 4.5 proves that there exists an estimator achieving the minimax rate $s\log(ep/s)/N$ for the ℓ_2-estimation risk (to the square) with absolutely no assumption on the output Y. In the case where a statistical model $Y = \mathrm{sign}(\langle X, t^* \rangle + \xi)$ holds, where ξ is independent of X then Theorem 4.5 shows that the RERM with logistic loss and SLOPE regularization achieves the minimax rate $s\log(ep/s)/N$ under no assumption on the noise ξ. In particular, ξ does not need to have any moment and, for instance, the mimimax rate $s\log(ep/s)/N$ can still be achieved when the noise has a Cauchy distribution. Moreover, this estimation rate holds with exponentially large probability as if the noise had a Gaussian distribution (cf. [Lecué and Mendelson, 2013]). This is a remarkable feature of Lipschitz loss functions genuinely understood in Huber's seminal paper [Huber, 1964].

	LASSO	SLOPE
$w(B)$	$\sqrt{\log p}$	1
ρ^*	$\frac{s}{\sqrt{N}}\sqrt{\log p}$	$\frac{s}{\sqrt{N}}\log\frac{ep}{s}$
$r(\rho^*)$	$\frac{s}{N}\log p$	$\frac{s}{N}\log\frac{ep}{s}$

Table 4.1 – Comparison of the key quantities involved in our study for the ℓ_1 (LASSO) and SLOPE norms

In Table 4.1, the different quantities playing an important role in our analysis have been collected for the ℓ_1 and SLOPE norms: the Gaussian mean width $w(B)$ of the unit ball B of the regularization norm, a radius ρ^* satisfying the sparsity equation and finally the L_2 estimation rate of convergence $r(\rho^*)$ summarizing the two quantities. As mentioned in Figure 4.1, having a large sub-differential at sparse vectors and a small Gaussian mean-width $w(B)$ is a good way to construct "sparsity inducing"regularization norms as it is, for instance the case of "atomic norms" (cf. [Chandrasekaran et al., 2012]).

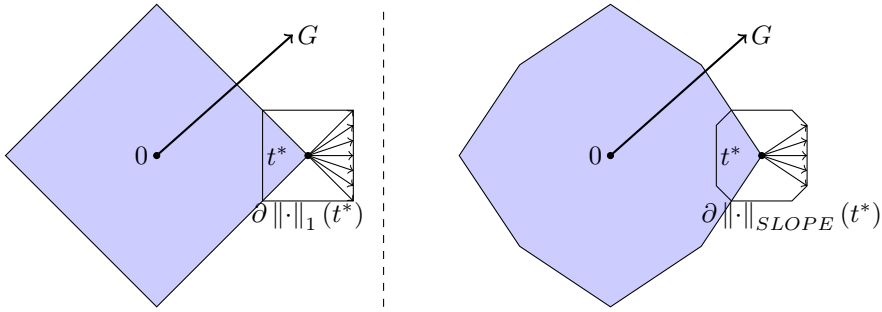


Figure 4.1 – **Gaussian complexity and size of the sub-differential for the ℓ_1 ans SLOPE norms:** A "large" sub-differential at sparse vectors and a small Gaussian mean width of the unit ball of the regularization norm is better for sparse recovery. In this figure, G represents a "typical" Gaussian vector used to compute the Gaussian mean width of the unit regularization norm ball.

4.4 Application to matrix completion via S_1-regularization

The second example involves matrix completion and uses the bounded setting from Section 4.2.6. The goal is to derive new results on two ways: the 1-bit matrix completion problem where entries are binary, and the quantile completion problem. The main theorems in this section yield upper bounds on completion in S_p norms ($1 \leq p \leq 2$) and on various excess risks. We also propose algorithms in order to compute efficiently the RERM in the matrix completion issue but with non differentiable loss and provide a simulation study. We first present a general theorem and then turn to specific loss functions because they induce a discussion about the Bernstein assumption and the κ parameter and lead to more particular theorems.

4.4.1 General result

In this section, we consider the matrix completion problem. Contrary to the introduction, we do not immediately focus on the case $Y \in \{-1, +1\}$. So for the moment, Y is a general real random variable and ℓ is any Lipschitz loss. The class is $F = \{\langle \cdot, M \rangle : M \in bB_\infty\}$,

where $bB_\infty = \{M = (M_{pq}) : \max_{p,q} |M_{pq}| \leq b\}$ and $b > 0$. As the design X takes its values in the canonical basis of $\mathbb{R}^{m\times T}$, the boundedness assumption is satisfied. Apart from that, the notations and assumptions are as in the introduction, that is, we assume that X satisfies Assumption 4.2, with parameters $(\underline{c}, \bar{c})$, and the penalty is the nuclear norm. Thus, the RERM is given by

$$\widehat{M} \in \operatorname*{arg\,min}_{M\in bB_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \ell(\langle X_i, M\rangle, Y_i) + \lambda \|M\|_{S_1} \right). \tag{4.16}$$

Statistical properties of (4.16) will follow from Theorem 4.3 since one can recast this problem in the setup of Section 4.2.6. The oracle matrix M^* is defined by $f^* = \langle \cdot, M^*\rangle$, that is, $M^* = \arg\min_{M\in bB_\infty} \mathbb{E}\ell(\langle M, X\rangle, Y)$.

Let us also introduce the matrix $\overline{M} = \arg\min_{M\in\mathbb{R}^{m\times T}} \mathbb{E}\ell(\langle M, X\rangle, Y)$. Note that $\langle \overline{M}, \cdot\rangle = \overline{f} = \arg\min_{f \text{ measurable}} \mathbb{E}\ell(f(X), Y)$. Our general results usually are on f^* rather than on $\overline{f}$ as it is usually impossible to provide rates on the estimation of $\overline{f}$ without stringent assumptions on Y and F. However, as noted in the introduction, in 1-bit matrix completion with the hinge loss, we have $\overline{M} = M^*$ without any extra assumption when $b = 1$ (this is a favorable case). On the other hand, to get fast rates in matrix completion with quantile loss requires that $\overline{M} = M^*$ (which is a stringent assumption in this setting).

Complexity function We first compute the complexity parameter $r(\cdot)$ as introduced in Definition 4.7. To that end one just needs to compute the global Rademacher complexity of the unit ball of the regularization function which is $B_{S_1} = \{A \in \mathbb{R}^{m\times T} : \|A\|_{S_1} \leq 1\}$:

$$\begin{aligned} \mathrm{Rad}(B_{S_1}) &= \mathbb{E} \sup_{\|A\|_{S_1}\leq 1} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i \langle X_i, A\rangle \right| = \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i X_i \right\|_{S_\infty} \\ &\leq c_0(\underline{c}, \bar{c}) \sqrt{\frac{\log(m+T)}{\min(m,T)}} \end{aligned} \tag{4.17}$$

where $\|\cdot\|_{S_\infty}$ is the operator norm (i.e. the largest singular value), the last inequality follows from Lemma 1 in [Koltchinskii et al., 2011] and $c_0(\underline{c}, \bar{c}) > 0$ is some constant that depends only on $\underline{c}$ and $\bar{c}$.

The complexity parameter $r(\cdot)$ is derived from Definition 4.7: for any $\rho \geq 0$,

$$r(\rho) = \left[\frac{CA\rho \mathrm{Rad}(B_{S_1})}{\sqrt{N}}\right]^{\frac{1}{2\kappa}} = \mathbf{C}\left[\rho\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right]^{\frac{1}{2\kappa}} \tag{4.18}$$

where from now the constants $\mathbf{C}$ depend only on $\underline{c}$, $\bar{c}$, b, A and κ.

Sparsity parameter The next important quantity is the sparsity parameter. Its expression in this particular case is, for any $\rho > 0$,

$$\Delta(\rho) = \inf\left\{\sup_{G\in\Gamma_{M^*}(\rho)} \langle H, G\rangle : H \in \rho S_{S_1} \cap ((\sqrt{mT}/\underline{c})r(2\rho))B_{S_2}\right\}$$

where $\Gamma_{M^*}(\rho)$ is the union of all the sub-differential of $\|\cdot\|_{S_1}$ in a S_1-ball of radius $\rho/20$ centered in M^*. Note that the normalization factor $\sqrt{mT}$ in the localization $(\sqrt{mT}r(2\rho))B_{S_2}$ comes from the "non normalized isotropic" property of X: $\underline{c}\,\|M\|_{S_2}^2/(mT) \leq \mathbb{E}\,\langle X, M\rangle^2 \leq \bar{c}\,\|M\|_{S_2}^2/(mT)$ for all $M \in \mathbb{R}^{m\times T}$. Now, we use a result from [Lecué and Mendelson, 2015b] to find a solution to the sparsity equation.

Lemma 4.2 (Lemma 4.4 in [Lecué and Mendelson, 2015b])**.** *There exists an absolute constant $c_1 > 0$ for which the following holds. If there exists $V \in M^* + (\rho/20)B_{S_1}$ such that* $\mathrm{rank}(V) \leq \left(c_1\rho/(\sqrt{mT}r(\rho))\right)^2$ *then* $\Delta(\rho) \geq 4\rho/5$.

It follows from Lemma 4.2 that the sparsity equation (4.8) is satisfied by ρ^* when it exists $V \in M^* + (\rho^*/20)B_{S_1}$ such that $\mathrm{rank}(V) = c_1\left(\rho^*/(\sqrt{mT}r(\rho^*))\right)^2$. Note obviously that V can be M^* itself, in this case, ρ^* can be taken such that $\mathrm{rank}(M^*) = c_1\left(\rho^*/(\sqrt{mT}r(\rho^*))\right)^2$. However, when M^* is not low-rank, it might still be that a low-rank approximation V of M^* is close enough to M^* w.r.t. the S_1-norm. As a consequence, if for some $s \in \{1, \ldots, \min(m,T)\}$ there exists a matrix V with rank at most s in $M^* + (\rho_s^*/20)B_{S_1}$ where

$$\rho_s^* = \mathbf{C}\,(smT)^{\frac{\kappa}{2\kappa-1}}\left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{2(2\kappa-1)}}. \tag{4.19}$$

then ρ_s^* satisfies the sparsity equation.

Following the remark at the end of Subsection 4.2.4, another possible choice is $\rho^* = 20\|M^*\|_{S_1}$ in order to get *norm dependent* rates. In the end, we choose $\rho^* = \mathbf{C}\min[\rho_s^*, \|M^*\|_{S_1}]$. We are now in a position to apply Theorem 4.3 to derive statistical properties for the RERM $\widehat{M}$ defined in (4.16).

Theorem 4.6. *Assume that Assumption 4.1, 4.2 and 4.3 hold. Consider the estimator in* (4.16) *with regularization parameter*

$$\lambda = \frac{c_0(\underline{c}, \bar{c})720}{7}\sqrt{\frac{\log(m+T)}{N\min(m,T)}} \tag{4.20}$$

where $c_0(\underline{c}, \bar{c})$ are the constants in Assumption 4.2. Let $s \in \{1, \dots, \min(m,T)\}$ and assume that there exists a matrix with rank at most s in $M^* + (\rho_s^*/20)B_{S_1}$. Then, with probability at least

$$1 - \mathbf{C}\exp\left(-\mathbf{C}s(m+T)\log(m+T)\right)$$

we have

$$\left\|\widehat{M} - M^*\right\|_{S_1} \leq \mathbf{C}\min\left\{(smT)^{\frac{\kappa}{2\kappa-1}}\left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{2(2\kappa-1)}}, \|M^*\|_{S_1}\right\},$$

$$\frac{1}{\sqrt{mT}}\left\|\widehat{M} - M^*\right\|_{S_2} \leq \mathbf{C}\min\left\{\left(\frac{s(m+T)\log(m+T)}{N}\right)^{\frac{1}{2(2\kappa-1)}}, \left(\|M^*\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right)^{\frac{1}{2\kappa}}\right\}$$

$$\mathcal{E}(\widehat{M}) \leq \mathbf{C}\min\left\{\left(\frac{s(m+T)\log(m+T)}{N}\right)^{\frac{\kappa}{2\kappa-1}}, \|M^*\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right\}.$$

Note that the interpolation inequality also allows to get a bound for the S_p norm, when $1 \leq p \leq 2$:

$$\frac{1}{(mT)^{\frac{1}{p}}}\left\|\widehat{M} - M^*\right\|_{S_p} \leq \mathbf{C}\min\left\{\left[\left(\frac{s^{2(p-1)+\kappa(2-p)}(m+T)^{p-1}}{\min(m,T)^{\frac{2-p}{2}}}\right)^{\frac{1}{p}}\sqrt{\frac{\log(m+T)}{N}}\right]^{\frac{1}{2\kappa-1}},\right.$$
$$\left.\|M^*\|_{S_1}^{\frac{p-1+\kappa(2-p)}{p\kappa}}\left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{p-1}{2\kappa p}}\left(\frac{1}{mT}\right)^{\frac{2-p}{p}}\right\}.$$

Theorem 4.6 shows that the sparsity dependent error rate in the excess risk bound is (for $s = \operatorname{rank}(M^*)$)

$$\left(\frac{\operatorname{rank}(M^*)(m+T)\log(m+T)}{N}\right)^{\frac{\kappa}{2\kappa-1}}$$

which is the classic excess risk bound under the margin assumption up to a log factor (cf. [Audibert and Tsybakov, 2007]). As for the S_2-estimation error, when $\kappa = 1$, we recover the classic S_2-estimation rate

$$\sqrt{\frac{\mathrm{rank}(M^*)(m+T)\log(m+T)}{N}}$$

which is minimax in general (up to log terms, e.g. take the quadratic loss when Y is bounded and compare to [Rohde and Tsybakov, 2011]).

4.4.2 Algorithm and Simulation Outlines

Since this part provides new methods and results on matrix completion, we propose an algorithm in order to compute efficiently the RERM using the hinge loss and the quantile loss. This section explains the structure of the algorithm that is used with specific loss functions in next sections. Although many algorithms exist for the least squares matrix completion, at our knowledge many of them treat only the exact recovery such as in [Cai et al., 2010] and [Mazumder et al., 2010], or at least they all deal with differentiable loss functions, see [Hsieh and Olsen, 2014]. On the other hand, the two losses that we mainly consider here are non differentiable because they are piecewise linear (in the case of hinge and $0,5$-quantile loss functions): new algorithms are hence needed. It has been often noted that the RERM with respect to the hinge loss or 0.5-quantile loss can been solved by a semidefinite programming but the cost is prohibitive for large matrices, say dimensions larger than 100. It actually works for small matrices as we ran SDP solver in Python in very small examples.

We propose here an *alternating direction method of multiplier* (ADMM) algorithm. For a clear and self-contained introduction to this class of algorithms, the reader is referred to [Boyd et al., 2011] and we do not explain all the details here and we keep the same vocabulary. When the optimization problem is a sum of two parts, the core idea is to split the problem by introducing an extra variable. In our case, the two following problems are equivalent:

$$\underset{M}{\text{minimize}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \ell(\langle X_i, M\rangle, Y_i) + \lambda \|M\|_{S_1} \right\}, \quad \underset{M,L}{\text{minimize}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \ell(\langle X_i, M\rangle, Y_i) + \lambda \|L\|_{S_1} \right\}$$
$$\text{subject to } M = L$$

Below, we use the scaled form and the $m \times T$ matrix U is then called the *scaled dual variable*. Note that the S_2 norm is also the Froebenius

norm and is thus elementwise. We can now exhibit the *augmented Lagrangian*:

$$L_\alpha(M,L,U) = \frac{1}{N}\sum_{i=1}^{N} \ell(\langle X_i, M\rangle, Y_i) + \lambda \|L\|_{S_1} + \frac{\alpha}{2}\|M-L+U\|_{S_2}^2 - \frac{\alpha}{2}\|U\|_{S_2}^2,$$

where α is a positive constant, called the *augmented Lagrange parameter*. The ADMM algorithm [Boyd et al., 2011] is then:

$$M^{k+1} = \underset{M}{\operatorname{argmin}} \left(\frac{1}{N}\sum_{i=1}^{N} \ell(\langle X_i, M\rangle, Y_i) + \frac{\alpha}{2}\left\|M - L^k + U^k\right\|_{S_2}^2 \right) \tag{4.21}$$

$$L^{k+1} = \underset{L}{\operatorname{argmin}} \left(\lambda \|L\|_{S_1} + \frac{\alpha}{2}\left\|M^{k+1} - L + U^k\right\|_{S_2}^2 \right) \tag{4.22}$$

$$U^{k+1} = U^k + M^{k+1} - L^{k+1}$$

The starting point (M^0, L^0, U^0) uses one random matrix with independent Gaussian entries for M^0 and two zero matrices for L^0 and U^0. Another choice of starting point is to use a previous estimator with a larger λ. The stopping criterion is, as explained in [Boyd et al., 2011], $\left\|M^{k+1} - M^k\right\|_{S_2}^2 + \left\|U^{k+1} - U^k\right\|_{S_2}^2 \leq \varepsilon$ for a fixed threshold ε. It means that it stops when both (U_k) and (M_k) start converging.

General considerations The second step (4.22) is independent of the loss function. It is well-known that the solution of this problem is $S_{\lambda/\alpha}(M^{k+1} + U^k)$ when $S_a(M)$ is the soft-thresholding operator with magnitude a applied to the singular values of the matrix M. It is defined for a rank r matrix M with SVD $M = U\Sigma V^\top$ where $\Sigma = \operatorname{diag}\left((d_i)_{1\leq i\leq r}\right)$ by $S_a(M) = US_a(\Sigma)V^\top$ where $S_a(\Sigma) = \operatorname{diag}\left((\max(0, d_i - a))_{1\leq i\leq r}\right)$.

It requires the SVD of a $m \times T$ matrix at each iteration and is the main bottleneck of this algorithm (the other main step (4.21) can be performed elementwise since the X_i's take their values in the canonical basis of $\mathbb{R}^{m\times T}$; so it needs only at most N operations). Two methods may be used in order to speed up the algorithm: efficient algorithms for computing the n largest singular values and the associate subspaces, such as the well-known PROPACK routine in Fortran. It can be plugged in order to solve (4.22) by computing the n largest and stop at this stage if the lowest computed singular values is lower than the threshold. It is

obviously more relevant when the target is expected to have a very small rank. This method has been implemented in Python and works well in practice even though the parameter n has to be tuned carefully. An alternative method is to use approximate SVD such as in [Halko et al., 2011].

Moreover, the first step (4.21) (which may be performed element-wise) has a closed form solution for hinge and quantile loss: it is a soft-thresholding applied to a specified quantity.

Simulated observations as well as real-world data (cf. the MovieLens dataset [2]) are considered in the examples below. Finally note that parameter λ is tuned by cross-validation.

4.4.3 1-bit matrix completion

In this subsection we assume that $Y \in \{-1, +1\}$, and we challenge two loss functions: the logistic loss, and the hinge loss. It is worth noting that the minimizer $\overline{M} = \arg\min_{M \in \mathbb{R}^{m \times T}} \mathbb{E}\ell(\langle M, X\rangle, Y)$ is not the same for both losses. For the hinge loss, it is known that it is the matrix formed by the Bayes classifier. This matrix has entries bounded by 1 so $M^* = \overline{M}$ as soon as $b = 1$. In opposite to this case, the logistic loss leads to a matrix $\overline{M}$ with entries formed by the odds ratio. It may even be infinite when there is no noise.

Logistic loss. Let us start by assuming that ℓ is the logistic loss. Thanks to Proposition 4.1 we know that $\kappa = 1$ for any b (A is also known, $A = 4\exp(2b)$) and therefore next result follows from Theorem 4.6. Note that we do not assume that $\overline{M}$ is in F and therefore our results provides estimation and prediction bounds for the oracle M^*.

Theorem 4.7 (1-bit Matrix Completion with logistic loss)**.** *Assume that Assumption 4.2 holds. Let $s \in \{1, \dots, \min(m, T)\}$ and assume that there exists a matrix with rank at most s in $M^* + (\rho_s^*/20)B_{S_1}$ where ρ_s^* is defined in* (4.19)*. With probability at least*

$$1 - \mathbf{C}\exp\left(-\mathbf{C}s\max(m, T)\log(m + T)\right)$$

the estimator

$$\widehat{M} \in \operatorname*{arg\,min}_{M \in bB_\infty} \left(\frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \exp\left(-Y_i \langle X_i, M\rangle\right)\right) + \lambda \|M\|_{S_1} \right) \quad (4.23)$$

2. available in http://grouplens.org/datasets/movielens/

with λ as in Equation (4.20) *satisfies*

$$\frac{1}{mT}\left\|\widehat{M}-M^*\right\|_{S_1} \leq \mathbf{C}\min\left\{s\sqrt{\frac{\log(m+T)}{N\min(m,T)}}, \frac{\|M^*\|_{S_1}}{mT}\right\},$$

$$\frac{1}{\sqrt{mT}}\left\|\widehat{M}-M^*\right\|_{S_2} \leq \mathbf{C}\min\left\{\sqrt{\frac{s\max(m,T)\log(m+T)}{N}}, \|M^*\|_{S_1}^{\frac{1}{2}}\left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{4}}\right\}$$

$$\mathcal{E}_{logistic}(\widehat{M}) \leq \mathbf{C}\min\left\{\frac{s\max(m,T)\log(m+T)}{N}, \|M^*\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right\}.$$

Using an interpolation inequality, it is easy to derive estimation bound in S_p for all $1 \leq p \leq 2$ as in Theorem 4.6 so we do not reproduce it here. Also, note that our bound on $\left\|\widehat{M}-M^*\right\|_{S_2}$ is of the same order as the one in [Lafond et al., 2014]. We actually now prove that this rate is minimax-optimal (up to log terms).

Theorem 4.8 (Lower bound with logistic loss)**.** *For a given matrix $M \in B_\infty$, define $\mathbb{P}_M^{\otimes N}$ as the probability distribution of the N-uplet $(X_i, Y_i)_{i=1}^N$ of i.i.d. pairs distributed like (X, Y) such that X is uniformly distributed on the canonical basis $(E_{p,q})$ of $\mathbb{R}^{m\times T}$ and $\mathbb{P}_M(Y = 1|X = E_{p,q}) = \exp(M_{pq})/[1+\exp(M_{pq})]$ for every $(p,q) \in \{1,\dots,m\}\times\{1,\dots,T\}$. Fix $s \in \{1,\dots,\min(m,T)\}$ and assume that $N \geq s(m+T)\log(2)/(8b^2)$. Then*

$$\inf_{\widehat{M}} \sup_{\substack{M^* \in bB_\infty \\ \mathrm{rank}(M^*) \leq s}} \mathbb{P}_{M^*}^{\otimes N}\left(\frac{1}{\sqrt{mT}}\left\|\widehat{M}-M^*\right\|_{S_2} \geq c\sqrt{\frac{(m+T)s}{N}}\right) \geq \beta$$

for some universal constants $\beta, c > 0$.

Also, as pointed out in the introduction, the quantity of interest is not the logistic excess risk, but the classification excess risk: let us remind that $R_{0/1}(M) = \mathbb{P}[(Y \neq \mathrm{sign}(\langle M, X\rangle)]$ for all $M \in \mathbb{R}^{m\times T}$. Even if we assume that $M^* = \overline{M}$, all that can be deduced from Theorem 2.1 in [Zhang, 2004] is that

$$\begin{aligned}\mathcal{E}_{0/1}(\widehat{M}) &= R_{0/1}(\widehat{M}) - \inf_{M\in\mathbb{R}^{m\times T}} R_{0/1}(M) \\ &\leq \mathbf{C}\sqrt{\mathcal{E}_{logistic}(\widehat{M})} \leq \mathbf{C}\sqrt{\frac{\mathrm{rank}(\overline{M})(m+T)\log(m+T)}{N}}.\end{aligned}$$

But this rate on the excess 0/1-risk may be much better under the margin assumption [Mammen and Tsybakov, 1999, Tsybakov, 2004] (cf. Equation (4.36) below). This motivates the use of the hinge loss instead of the logistic loss, for which the results in [Zhang, 2004] do not lead to a loss of a square root in the rate.

Hinge loss. As explained above, the choice $b = 1$ ensures $\overline{M} = M^*$ without additional assumption. Thanks to Proposition 4.3 we know that as soon as $\inf_{p,q} |\overline{M}_{p,q} - 1/2| \geq \tau$ for some $\tau > 0$, the Bernstein assumption is satisfied by the hinge loss with $\kappa = 1$ and $A = 1/(2\tau)$. This assumption seems very mild in many situations and we derive the results with it.

Theorem 4.9 (1-bit Matrix Completion with hinge loss)**.** *Assume that Assumption 4.2 holds. Assume that* $\inf_{p,q} |P(Y = 1|X = E_{p,q}) - 1/2| \geq \tau$ *for some* $\tau > 0$*. Let* $s \in \{1, \dots, \min(m, T)\}$ *and assume that there exists a matrix with rank at most* s *in* $\overline{M} + (\rho_s^*/20)B_{S_1}$ *where* ρ_s^* *is defined in* (4.19)*. With probability at least*

$$1 - \mathbf{C} \exp\left(-\mathbf{C} s \max(m, T) \log(m + T)\right)$$

the estimator

$$\widehat{M} \in \operatorname*{arg\,min}_{M \in B_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \left(1 - Y_i \langle X_i, M \rangle\right)_+ + \lambda \|M\|_{S_1} \right) \tag{4.24}$$

with λ *as in Equation* (4.20) *satisfies*

$$\frac{1}{mT} \left\| \widehat{M} - \overline{M} \right\|_{S_1} \leq \mathbf{C} \min \left\{ s \sqrt{\frac{\log(m+T)}{N \min(m,T)}}, \frac{\left\|\overline{M}\right\|_{S_1}}{mT} \right\},$$

$$\frac{1}{\sqrt{mT}} \left\| \widehat{M} - \overline{M} \right\|_{S_2} \leq \mathbf{C} \min \left\{ \sqrt{\frac{s(m+T)\log(m+T)}{N}}, \left\|\overline{M}\right\|_{S_1}^{\frac{1}{2}} \left(\frac{\log(m+T)}{N \min(m,T)} \right)^{\frac{1}{4}} \right\}$$

$$\mathcal{E}_{hinge}(\widehat{M}) \leq \mathbf{C} \min \left\{ \frac{s(m+T)\log(m+T)}{N}, \left\|\overline{M}\right\|_{S_1} \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \right\}.$$

In this case, [Zhang, 2004] implies that the excess risk bound for the classification error (using the 0/1-loss) is the same as the one for the hinge loss: it is therefore of the order of $\operatorname{rank}(\overline{M}) \max(m, T)/N$.

Note that the rate $\operatorname{rank}(\overline{M}) \max(m, T)/N$ for the classification excess error was only reached in [Cottet and Alquier, 2016] up to our knowledge (using the PAC-Bayesian technique from [Catoni, 2004, 2007, Mai

and Alquier, 2015, Alquier et al., 2016]), in the very restrictive noiseless setting - that is, $P(Y = 1|X = E_{p,q}) \in \{0, 1\}$ which is equivalent to $P(Y = \text{sign}(\langle \overline{M}, X \rangle) = 1$. Here this rate is proved to hold in the general case. Other works, including [Srebro et al., 2005], obtained only rates in $1/\sqrt{N}$. Finally, we prove that this rate is the minimax rate in the next result.

Theorem 4.10 (Lower bound with hinge loss)**.** *For a given matrix $M \in B_\infty$, let $\mathbb{E}_M^{\otimes N}$ be the expectation w.r.t. the N-uplet $(X_i, Y_i)_{i=1}^N$ of i.i.d. pairs distributed like (X, Y) such that X is uniformly distributed on the canonical basis $(E_{p,q})$ of $\mathbb{R}^{m\times T}$ and $\mathbb{P}_M(Y = 1|X = E_{p,q}) = M_{pq}$ for every $(p, q) \in \{1, \dots, m\} \times \{1, \dots, T\}$. Fix $s \in \{1, \dots, \min(m, T)\}$ and assume that $N \geq s \max(m, T) \log(2)/8$. Then*

$$\inf_{\widehat{M}} \sup_{\substack{M^* \in B_\infty \\ \text{rank}(M^*) \leq r}} \mathbb{E}_{M^*}^{\otimes N} \left(\mathcal{E}_{hinge}(\widehat{M}) \right) \geq c \frac{s \max(m, T)}{N}$$

for some universal constants $c > 0$.

Theorem 4.10 provides a minimax lower bound in expectation whereas Theorem 4.9 provides an excess risk bound with large deviation. The two residual terms of the excess hinge risk from Theorem 4.10 and Theorem 4.9 match up to the $\log(m + T)$ factor.

Simulation Study. As the hinge loss has not been often studied in the matrix context, we provide many simulations in order to show the robustness of our method and the opportunity of using the hinge loss rather than the logistic loss. We follow the simulations ran in [Cottet and Alquier, 2016] and compare several methods. An estimator based on the logistic model, studied in [Davenport et al., 2014], is also challenged [3].

A first set of simulations. The simulations are all based on a low-rank 200×200 matrix $M^\star$ from which the data are generated and which is the target for the predictions. $M^\star$ is also a minimizer of $R_{0/1}$ so the error criterion that we will report for a matrix M is the difference of the predictions between $M^\star$ and M, which is $\mathbb{P}[\text{sign}(\langle M^\star, X \rangle) \neq \text{sign}(\langle M, X \rangle)]$.

3. In the followings, the four estimators will be referred to *Hinge* for estimator given in (4.24), *Hinge Bayes* and *Logit Bayes* for the two Bayesian estimators from [Cottet and Alquier, 2016] with respectively hinge and logistic loss functions, and *Logit* for the estimator from [Davenport et al., 2014]. The Bayesian estimators use the Gammma prior distribution.

The X_i's correspond to 20% of the entries randomly picked so the misclassification rate is also $1/mT \sum_{p,q} I\{\text{sign}(M_{p,q}) \neq \text{sign}(M^\star_{p,q})\}$.

Two different scenarios are tested: the first one (called A), involves a matrix $M^\star$ with only entries in $\{-1, +1\}$ so the Bayes classifier is low rank and favors the hinge loss. The second test (called B) involves a matrix $M^\star = LR^\top$ where L, R have i.i.d. Gaussian entries and the rank is the number of columns. In this case, the Bayes matrix contains the signs of a low-rank matrix, but it is not itself low rank in general. We also test the impact of the noise structure on the results:

1. (noiseless) $Y_i = \text{sign}(\langle M^\star, X_i \rangle)$
2. (logistic) $Y_i = \text{sign}(\langle M^\star, X_i \rangle + Z_i)$, where Z_i follows a logistic distribution
3. (switch) $Y_i = \epsilon_i \text{sign}(\langle M^\star, X_i \rangle)$ where $\epsilon_i = (1-p)\delta_1 + p\delta_{-1}$

Finally, we run all the simulations on rank 3 and rank 5 matrices. λ is tuned by cross validation. All the simulations are run one time.

	Model	A1	A2 ($p = .1$)	A3	B1	B2 ($p = .1$)	B3
Rank 3	Hinge	0	0	14.5	6.7	10.9	21.0
	Logit	0	0.5	17.3	5.1	10.7	19.8
	Hinge Bayes	0	0.1	8.5	5.3	10.8	22.1
	Logit Bayes	0	0.5	16.0	4.1	10.1	16.0
Rank 5	Hinge	0	0.8	29.0	11.7	19.3	23.3
	Logit	0	3.1	30.1	9.0	18.3	22.1
	Hinge Bayes	0	0.5	27	9.4	17.9	24.4
	Logit Bayes	0	4.4	32.5	7.8	17.3	21.5

Table 4.2 – Misclassification error rates on simulated matrices in various cases. Model $\in \{A, B\}\{1, 2, 3\}$ refers to scenario $\in \{A, B\}$ and noise structure $\in \{1, 2, 3\}$. For the noise-free Model $= A0$, the 0 column shows the exact reconstruction property of all procedures.

The results are very similar among the methods, see Table 4.2. The logistic loss performs better for matrices of type B and especially for high level of noise in the logistic data generation as expected. For type A matrices, the hinge loss performs slightly better. The Bayesian models performs as good as the frequentist estimators even though the program solved is not convex.

Impact of the noise level. The second experiment is a focus on the switch noise and matrices that are well separated (as A2 in the previous example). The noise lies between $p = 0$ and almost full noise ($p = .4$).

The performance of the RERM with the hinge loss is slightly worse than the Bayesian estimator with hinge loss but always better than the RERM with the logistic loss, see Figure 4.4.3.

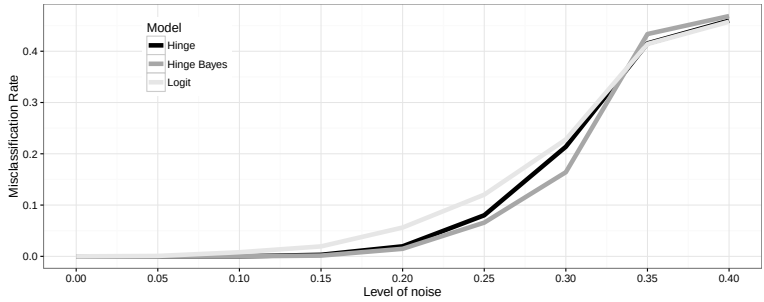


Figure 4.2 – Misclassification error rates for a large range of switch noise (noise structure number 3).

Real dataset. We finally run the hinge loss estimator on the MovieLens dataset. The ratings, that lie in $\{1, 2, 3, 4, 5\}$, are split between good ratings $(4, 5)$ and bad ratings (others). The goal is therefore to predict whether the user will like a movie or not. On a test set that contains 20% of the data, the misclassification rate in prediction are almost the same for all the methods (Table 4.3).

Model	Hinge Bayes	Logit	Hinge
misclassification rate	.28	.27	.28

Table 4.3 – Misclassification Rate on MovieLens 100K dataset

4.4.4 Quantile loss and median matrix completion

The matrix completion problem with continuous entries has almost always been tackled with a penalized least squares estimator [Candès and Plan, 2010, Koltchinskii et al., 2011, Klopp, 2014, Lecué and Mendelson, 2015b, Mai and Alquier, 2015], but the use of other loss functions may be very interesting in this case too. Our last result on matrix completion is a result for the quantile loss ρ_τ for $\tau \in (0, 1)$. Let us recall that $\rho_\tau(u) = u(\tau - I(u \leq 0))$ for all $u \in \mathbb{R}$ and $\ell_M(x, y) = \rho_\tau(y - \langle M, x \rangle)$. While the

aforementionned references provided ways to estimate the conditional mean of $Y|X = E_{p,q}$, here, we thus provide a way to estimate conditional quantiles of order τ. When $\tau = 0.5$, it actually estimates the conditional median, which is known to be an indicator of central tendency that is more robust than the mean in the presence of outliers. On the other hand, for large and small τ's (for example the 0.05 and 0.95 quantiles), this allows to build confidence intervals for $Y|X = E_{p,q}$. Confidence bounds for the entries of matrices in matrix completion problems are something new up to our knowledge.

The following result studies a particular case in which the Bernstein Assumption is proved in Proposition 4.4. Following [van de Geer, 2016], it assumes that the conditional distribution of Y given X is continuous and that the density is not too small on the domain of interest – this ensures that Bernstein's condition is satisfied with $\kappa = 1$ and A depending on the lower bound on the density, see Section 4.6 for more details. It can easily be derived for a specific distribution such as Gaussian, Student and even Cauchy. But we also have to assume that $\overline{M} \in bB_\infty$, or in other words $\overline{M} = M^*$, which is a more stringent assumption: in practice, it meands that we should know *a priori* an upper bound b on the quantiles to be estimated.

Theorem 4.11 (Quantile matrix completion)**.** *Assume that Assumption 4.2 holds. Let $b > 0$ and assume that $\overline{M} \in bB_\infty$. Assume that for any (p, q), $Y|X = E_{p,q}$ has a density with respect to the Lebesgue measure, g, and that $g(u) > 1/c$ for some constant $c > 0$ for any u such that $|u - \overline{M}_{i,j}| \leq 2b$. Let $s \in \{1, \dots, \min(m, T)\}$ and assume that there exists a matrix with rank at most s in $\overline{M} + (\rho_s^*/20)B_{S_1}$ where ρ_s^* is defined in* (4.19)*. Then, with probability at least*

$$1 - \mathbf{C} \exp\left(-\mathbf{C} s \max(m, T) \log(m + T)\right)$$

the estimator

$$\widehat{M} \in \operatorname*{arg\,min}_{M \in bB_\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \rho_\tau(Y_i - \langle X_i, M \rangle) + \lambda \|M\|_{S_1} \right) \tag{4.25}$$

with $\lambda = c_0(\underline{c}, \bar{c}) \sqrt{\log(m + T)/(N \min(m, T))}$ satisfies

$$\frac{1}{mT} \left\| \widehat{M} - \overline{M} \right\|_{S_1} \leq \mathbf{C} \min \left\{ s \sqrt{\frac{\log(m + T)}{N \min(m, T)}}, \frac{\left\| \overline{M} \right\|_{S_1}}{mT} \right\},$$

$$\frac{1}{\sqrt{mT}}\left\|\widehat{M}-\overline{M}\right\|_{S_2} \leq \mathbf{C}\min\left\{\sqrt{\frac{s(m+T)\log(m+T)}{N}}, \left\|\overline{M}\right\|_{S_1}^{\frac{1}{2}}\left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{4}}\right\}$$

$$\mathcal{E}_{quantile}(\widehat{M}) \leq \mathbf{C}\min\left\{\frac{s(m+T)\log(m+T)}{N}, \left\|\overline{M}\right\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right\}.$$

We obtain the same rate as for the penalized least squares estimator that is $\sqrt{s(m+T)\log(m+T)/N}$ (cf. [Rohde and Tsybakov, 2011, Koltchinskii et al., 2011]).

Simulation study.

The goal of this part is to challenge the regularized least squares estimator by the RERM with quantile loss. The quantile used here is therefore the median. The main conclusion of our study is that median based estimators are more robust to outliers and noise than mean based estimators. We first test them on simulated datasets and then turn to use a real dataset.

Simulated matrices. The observations come from a base matrix $M^\star$ which is a 200×200 low rank matrix. It is built by $M^\star = LR^\top$ where the entries of L, R are i.i.d. gaussian and L, R have 3 columns (and therefore, the rank of $M^\star$ is 3). The X_i's correspond to 20% randomly picked entries. The criterion that we retain is the l_1 reconstruction of $M^\star$ that is: $1/mT\sum_{p,q}|M^\star_{p,q} - M_{p,q}|$.

The observations are made according to this flexible model:

$$Y_i = \langle M^\star, X\rangle + z_i + o\zeta_i.$$

z_i is the noise, o is the magnitude of outliers and ζ_i is the outlier indicator parametrized by the share p such that $\zeta_i = p/2\delta_{-1} + (1-p)\delta_0 + p/2\delta_1$. The different parameters for the different scenarios are summarized in Table 4.4.

On the first experiment, p is fixed to 10% and the magnitude o increases. As expected for least squares, the results are better for low magnitude of outliers (it corresponds to the penalized maximum likelihood estimator), see Figure 4.3. Quickly, the performance of the least squares estimator is getting worse and when the outliers are large enough, the best least squares predictor is a matrix with null entries. In opposite to this estimator, the median of the distribution is almost not affected by outliers and it is completely in line with the results: the performances

are strictly the same for mid-range to high-range magnitude of outliers. The robustness of the quantile reconstruction is totally independent to the magnitude of the outliers.

	z_i	o	ζ_i
Figure 4.3	$\mathcal{N}(0, 1/4)$	$o = 0..30$	$p = 0.1$
Figure 4.4	$\mathcal{N}(0, 1/4)$	10	$p = 0..0.25$
Figure 4.5	$t_\alpha, \alpha = 1..10$	0	$p = 0$

t_α: t-distribution with α degrees of freedom.

Table 4.4 – Parameters and distributions of the simulations

A second experiment involves fixed magnitude of outliers but the share of them increases, see Figure 4.4. The median completion is, as expected, more robust and the results deteriorate less than the ones from least squares. When the outliers ratio is greater than 20%, the least squares estimator completely fails while the median completion still works.

The third simulation involves non gaussian noise without outliers: we use the t-distribution, that has heavy tails. In this challenge, a lower degree of freedom involves heavier tails and the worst case is for Student distribution with degree 1. We can see that the least squares is inadequate for small degrees of freedom (1 to 2) and behaves better than the median completion for larger degrees of freedom, see Figure 4.5.

Real dataset. The last experiment involves the MovieLens dataset. We keep one fifth of the sample for test set to check the prediction accuracy. Even though the least squares estimator remains very efficient in the standard case, see Table 4.5, the results are quite similar for the MAE criterion. In a second step, we add artificial outliers. In order to do that, we change 20% of 5 ratings to 1 ratings. It can be seen as malicious users that change ratings in order to distort the perception of some movies. As expected, it depreciates the least squares estimator performance but the median estimator returns almost as good performances as in the standard case.

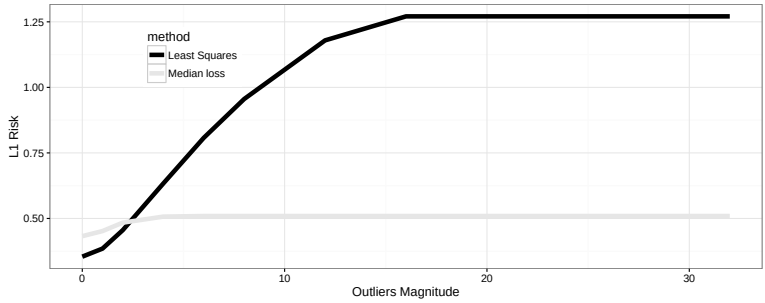


Figure 4.3 – l_1 reconstruction for different magnitude of outliers

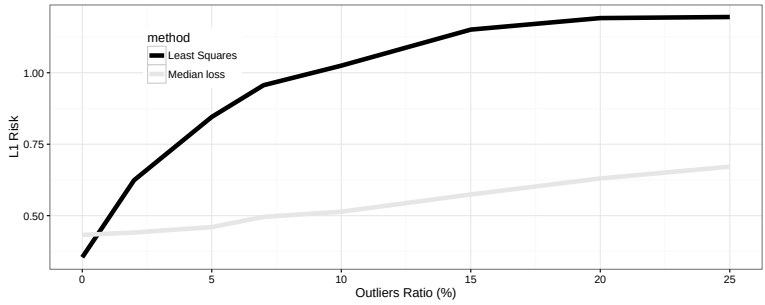


Figure 4.4 – l_1 reconstruction for different percentage of outliers

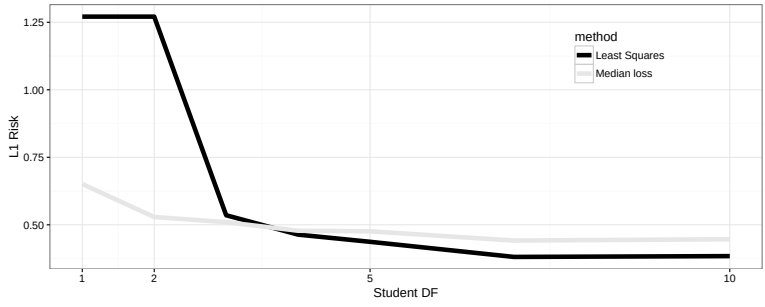


Figure 4.5 – l_1 reconstruction for student noise with various magnitude degrees of freedom

	MSE	MAE
Raw Data, LS	0.89	0.75
Raw Data, Median	0.93	0.75
Outliers, LS	1.04	0.84
Outliers, Median	0.96	0.78

Table 4.5 – Prediction power of Least Squares and Median Loss on MovieLens 100K dataset

4.5 Kernel methods via the hinge loss and a RKHS-norm regularization

In this section, we consider regularization methods in some general Reproducing Kernel Hilbert Space (RKHS) (cf. [Cucker and Smale, 2002], Chapter 4 in [Steinwart and Christmann, 2008] or Chapter 3 of [Vapnik, 1998] for general references on RKHS).

Unlike the previous examples, the regularization norm here, which is the norm $\|\cdot\|_{\mathcal{H}_K}$ of a RKHS $\mathcal{H}_K$, is not associated with some "hidden" concept of sparsity. In particular, RKHS norms have no singularity since they are differentiable at any point except in 0. As a consequence the sparsity parameter $\Delta(\rho)$ cannot be larger than $4\rho/5$, i.e. ρ does not satisfy the sparsity equation, unless the set $\Gamma_{f^*}(\rho)$ contains 0 that is for $\rho \geq 20 \|f^*\|_{\mathcal{H}_K}$. Indeed, one key observation is that any norm is non differentiable at 0 and that its subdifferential at 0 is somehow extremal:

$$\partial \|\cdot\| (0) = B_* := \{f : \|f\|_* \leq 1\}, \tag{4.26}$$

where $\|\cdot\|_*$ is the dual norm.

As a consequence, the rates obtained in this section do not depend on some *hidden sparsity parameter* associated with the oracle f^* but on the RKHS norm at f^*, that is $\|f^*\|_{\mathcal{H}_K}$. The aim of this section is therefore to show that our main results apply beyond "sparsity inducing regularization methods" by showing that "classic" regularization method, inducing smoothness for instance, may also be analyzed the same way and fall into the scope of Theorem 4.2 and Theorem 4.3. This section also shows an explicit expression for the Gaussian mean-width with localization as used in Definition 4.8 (a sharper way to measure statistical complexity via a local $r(\cdot)$ function provided below).

Mathematical background In this setup, the data are still N i.i.d. pairs $(X_i, Y_i)_{i=1}^N$ where the X_i's take their values in some set $\mathcal{X}$ and $Y_i \in \{-1, +1\}$. A "similarity measure" is provided over the set $\mathcal{X}$ by means of a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ so that $x_1, x_2 \in \mathcal{X}$ are "similar" when $K(x_1, x_2)$ is small. One can think for instance of $\mathcal{X}$ the set of all DNA sequences (that is finite words over the alphabet $\{A, T, C, G\}$) and $K(w_1, w_2)$ is the minimal number of changes like insertion, deletion and mutation needed to transform word $w_1 \in \mathcal{X}$ into word $w_2 \in \mathcal{X}$.

The core idea behind kernel methods is to transport the design data X_i's from $\mathcal{X}$ to a Hilbert space via the application $x \to K(x, \cdot)$ and then construct statistical procedures based on the "transported" dataset $(K(X_i, \cdot), Y_i)_{i=1}^N$. The advantage of doing so is that the space where the $K(X_i, \cdot)$'s belong have much structure than the initial set $\mathcal{X}$ which may have no algebraic structure at all. The first thing to set is to define somehow the "smallest" Hilbert space containing all the functions $x \to K(x, \cdot)$. We recall now one classic way of doing so that will be used later to define the objects that need to be considered in order to construct RERM in this setup and to obtain estimation rates for them via Theorem 4.2 and Theorem 4.3.

Recall that if $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel such that $\|K\|_{L_2} < \infty$, then by Mercer's theorem, there is an orthogonal basis $(\phi_i)_{i\in\mathbb{N}}$ of L_2 such that $\mu\otimes\mu$-almost surely, $K(x, x') = \sum_{i=1}^\infty \lambda_i \phi_i(x)\phi_i(x')$ where $(\lambda_i)_{i\in\mathbb{N}}$ is the sequence of eigenvalues of the positive self-adjoint integral operator T_K (arranged in a non-increasing order) defined for every $f \in L_2$ and μ-almost every $x \in \mathcal{X}$ by

$$(T_K f)(x) = \int K(x, x')f(x')d\mu(x').$$

In particular, for all $i \in \mathbb{N}$, ϕ_i is an eigenvector of T_K corresponding to the eigenvalue λ_i; and $(\phi_i)_i$ is an orthonormal system in L_2.

The reproducing kernel Hilbert space $\mathcal{H}_K$ is the set of all function series $\sum_{i=1}^\infty a_i K(x_i, \cdot)$ converging in L_2 endowed with the inner product

$$\left\langle \sum a_i K(x_i, \cdot), \sum b_j K(x'_j, \cdot) \right\rangle = \sum_{i,j} a_i b_j K(x_i, x'_j)$$

where a_i, b_j's are any real numbers and the x_i's and x'_j's are any points in $\mathcal{X}$.

Estimator. The RKHS $\mathcal{H}_K$ is therefore a class of functions from $\mathcal{X}$ to $\mathbb{R}$ that can be used as a learning model and the norm naturally associated

to its Hilbert structure can be used as a regularization function. Given a Lipschitz loss function ℓ, the oracle is defined as

$$f^* \in \operatorname*{arg\,min}_{f \in \mathcal{H}_K} \mathbb{E}\ell_f(X, Y)$$

and it is believed that $\|f^*\|_{\mathcal{H}_K}$ is small which justified the use of the RERM with regularization function given by the RKHS norm $\|\cdot\|_{\mathcal{H}_K}$:

$$\hat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{H}_K} \left(\frac{1}{N} \sum_{i=1}^{N} \ell_f(X_i, Y_i) + \lambda \|f\|_{\mathcal{H}_K} \right)$$

Statistical properties of this RERM may be obtained from Theorem 4.2 in the subgaussian case and from Theorem 4.3 in the bounded case. To that end, we only have to compute the Gaussian mean width and/or the Rademacher complexities of $B_{\mathcal{H}_K}$. In this example, we rather compute the localized version of those quantities because it is possible to derive explicit formula. They are obtained by intersecting the ball with $r\mathcal{E}$. In order not to induce any confusion, we still use the global ones in estimation bounds.

Localized complexity parameter. The goal is to compute $w(\rho B_{\mathcal{H}_K} \cap r\mathcal{E})$ and $\mathrm{Rad}(\rho B_{\mathcal{H}_K} \cap r\mathcal{E})$ for all $\rho, r > 0$ where $B_{\mathcal{H}_K} = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq 1\}$ is the unit ball of the RKHS and $\mathcal{E} = \{f \in \mathcal{H}_K : \mathbb{E}f(X)^2 \leq 1\}$ is the ellipsoid associated with X. In the following, we embed the two sets $B_{\mathcal{H}_K}$ and $\mathcal{E}$ in $l_2 = l_2(\mathbb{N})$ so that we simply have to compute the Gaussian mean width and the Rademacher complexities of the intersection of two ellipsoids sharing the same coordinates structure.

The unit ball of $\mathcal{H}_K$ can be constructed from the eigenvalue decomposition of T_K by considering the feature map $\Phi : \mathcal{X} \to l_2$ defined by $\Phi(x) = \left(\sqrt{\lambda_i}\phi_i(x)\right)_{i \in \mathbb{N}}$ and then the unit ball of $\mathcal{H}_K$ is just

$$B_{\mathcal{H}_K} = \left\{ f_\beta(\cdot) = \langle \beta, \Phi(\cdot) \rangle : \|\beta\|_{l_2} \leq 1 \right\}.$$

One can use the feature map Φ to show that there is an isometry between the two Hilbert spaces $\mathcal{H}_K$ and l_2 endowed with the norm $\|\beta\|_K = \left(\sum \beta_i^2/\lambda_i\right)^{1/2}$. The unit ball of l_2 endowed with the norm $\|\cdot\|_K$ is an ellipsoid denoted by $\mathcal{E}_K$.

Let us now determine the ellipsoid in l_2 associated with the design X obtained via this natural isomorphism $\beta \in l_2 \to f_\beta(\cdot) = \langle \beta, \Phi(\cdot) \rangle \in \mathcal{H}_K$

between l_2 and $\mathcal{H}_K$. Since $(\phi_i)_i$ is an orthonormal system in L_2, the covariance operator of $\Phi(X)$ in l_2 is simply the diagonal operator with diagonal elements $(\lambda_i)_i$. As a consequence the ellipsoid associated with X is isomorphic to $\tilde{\mathcal{E}} = \{\beta \in l_2 : \mathbb{E}\langle \beta, \Phi(X)\rangle^2 \leq 1\}$; it has the same coordinate structure as the canonical one in l_2 endowed with $\|\cdot\|_K$: $\tilde{\mathcal{E}} = \{\beta \in l_2 : \sum \lambda_i \beta_i^2 \leq 1\}$. So that, we obtain

$$w(K_\rho(f^*) \cap r\mathcal{E}_{f^*}) = w(\rho\mathcal{E}_K \cap r\tilde{\mathcal{E}}) \sim \left(\sum_j (\rho^2 \lambda_j) \wedge r^2 \right)^{1/2} \tag{4.27}$$

where the last inequality follows from Proposition 2.2.1 in [Talagrand, 2005] (note that we defined the Gaussian mean widths in Definition (4.4) depending on the covariance of X). We also get from Theorem 2.1 in [Mendelson, 2004] that

$$\mathrm{Rad}(K_\rho(f^*) \cap r\mathcal{E}_{f^*}) \sim \left(\sum_j (\rho^2 \lambda_j) \wedge r^2 \right)^{1/2}. \tag{4.28}$$

Note that unlike the previous examples, we do not have to assume isotropicity of the design. Indeed, in the RKHS case, the unit ball of the regularization function is isomorphic to the ellipsoid $\mathcal{E}_K$. Since $\mathcal{E}$ is also an ellipsoid having the same coordinates structure as $\mathcal{E}_K$ (cf. paragraph above), for all $\rho, r > 0$, the intersection $\rho B_{\mathcal{H}_K} \cap r\mathcal{E}$ is equivalent to an ellipsoid, meaning that, it contains an ellipsoid and is contained in a multiple of this ellipsoid. Therefore, the Gaussian mean width and the Rademacher complexity of $\rho B_{\mathcal{H}_K} \cap r\mathcal{E}$ has been computed without assuming isotropicity (thanks to general results on the complexity of Ellipsoids from Proposition 2.2.1 in [Talagrand, 2005] and Theorem 2.1 in [Mendelson, 2004]).

It follows from (4.27) and (4.28) that the Gaussian mean width and the Rademacher complexities are equal. Therefore, up to constant (L in the subgaussian case and b in the bounded case), the two subgaussian and bounded setups may be analyzed at the same time. Nevertheless, since we will only consider in this setting the hinge loss and that the Bernstein condition (cf. Assumption 4.3) with respect to the hinge loss has been studied in Proposition 4.3 only in the bounded case. We therefore continue the analysis only for the bounded framework.

We are now able to identify the complexity parameter of the problem. We actually do not use the localization in this and rather use only the

global complexity parameter as defined in Definition 4.7: for all $\rho > 0$:

$$r(\rho) = \left[\frac{\mathbf{C}\rho \left(\sum_j \lambda_j \right)^{1/2}}{\sqrt{N}} \right]^{\frac{1}{2\kappa}} \tag{4.29}$$

where $\kappa \geq 1$ is the Bernstein parameter.

Results in the bounded setting Finally, let us discuss about the boundedness assumption. It is known (cf., for instance, Lemma 4.23 in [Steinwart and Christmann, 2008]) that if the kernel K is bounded then the functions in the RKHS $\mathcal{H}_K$ are bounded: for any $f \in \mathcal{H}_K$, $\|f\|_{L_\infty} \leq \|K\|_\infty \|f\|_{\mathcal{H}_K}$ where $\|K\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}$. As a consequence, if one restricts the search space of the RERM to a RKHS ball of radius R, one has $F := RB_{\mathcal{H}_K} \subset \|K\|_\infty B_{L_\infty}$ and therefore the boundedness assumption is satisfied by F. However, note that a refinement of the proof of Theorem 4.14 using a boundedness parameter b depending on the radius of the RKHS balls used while performing the peeling device yields statistical properties for the RERM with no search space constraint. For the sake of shortness, we do not provide this analysis here.

We are now in a position to provide estimation and prediction results for the RERM

$$\hat{f} \in \operatorname*{arg\,min}_{f \in RB_{\mathcal{H}_K}} \left(\frac{1}{N} \sum_{i=1}^{N} (1 - Y_i f(X_i))_+ + \frac{\mathbf{C} \left(\sum_j \lambda_j \right)^{1/2}}{\sqrt{N}} \|f\|_{\mathcal{H}_K} \right) \tag{4.30}$$

where the choice of the regularization parameter λ follows from Theorem (4.3) and (4.28) (for $r = +\infty$). Note that unlike the examples in the previous sections, we do not have to find some radius ρ^* satisfying the sparsity equation (4.8) to apply Theorem 4.3 since we simply take $\rho^* = 20 \|f^*\|_{\mathcal{H}_K}$ to insure that $0 \in \Gamma_{f^*}(\rho^*)$.

Theorem 4.12. *Let $\mathcal{X}$ be some space, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded kernel and denote by $\mathcal{H}_K$ the associated RKHS. Denote by $(\lambda_i)_i$ the sequence of eigenvalues associated to $\mathcal{H}_K$ in L_2. Assume that the Bayes rule $\overline{f}$ from* (4.35) *belongs to $RB_{\mathcal{H}_K}$ and that the margin assumption* (4.36) *is satisfied for some $\kappa \geq 1$.*

Then the RERM defined in (4.30) *satisfies with probability larger than*

$$1 - \mathbf{C}\exp\left(-\mathbf{C}N^{1/2\kappa}\left(\left\|\overline{f}\right\|_{\mathcal{H}_K}\left(\sum_j \lambda_j\right)^{1/2}\right)^{(2\kappa-1)/\kappa}\right),$$

that

$$\left\|\hat{f} - \overline{f}\right\|_{L_2} \leq \mathbf{C}\left[\frac{\left\|\overline{f}\right\|_{\mathcal{H}_K}\left(\sum_j \lambda_j\right)^{1/2}}{\sqrt{N}}\right]^{1/2\kappa}$$

$$\text{and } \mathcal{E}_{hinge}(\hat{f}) \leq \mathbf{C}\frac{\left\|\overline{f}\right\|_{\mathcal{H}_K}\left(\sum_j \lambda_j\right)^{1/2}}{\sqrt{N}}$$

where $\mathcal{E}(\hat{f})$ *is the excess hinge risk of* $\hat{f}$*.*

Note that classic procedures in the literature on RKHS are mostly developed in the classification framework. They are usually based on the hinge loss and the regularization function is the square of the RKHS norm. For such procedures, oracle inequalities have been obtained in Chapter 7 from [Steinwart and Christmann, 2008] under the margin assumption (cf. [Tsybakov, 2004]). A result that is close to the one obtained in Theorem 4.12 is Corollary 4.12 in [Mendelson and Neeman, 2010]. Assuming that $\|Y\|_\infty \leq \mathbf{C}$, $\mathcal{X} \subset \mathbb{R}^d$, $\|K\|_\infty \leq 1$, that the eigenvalues of the integral operator satisfies

$$\lambda_i \leq ci^{-1/p} \tag{4.31}$$

for some $0 < p < 1$ and that the eigenvectors (ϕ_i) are such that $\|\phi_i\|_\infty \leq A$ for any i and some constant A then the RERM $\tilde{f}$ over the entire RKHS space, w.r.t. the quadratic loss and for a regularization function of the order of (up to logarithmic terms)

$$f \mapsto \rho(\|f\|_{\mathcal{H}}) := \max\left(\frac{\|f\|_{\mathcal{H}}^{2p/(1+p)}}{N^{1/(1+p)}}, \frac{\|f\|_{\mathcal{H}}^2}{N}\right) \tag{4.32}$$

satisfies with large probability an oracle inequality like

$$\mathbb{E}(Y - \tilde{f}(X))^2 \leq \inf_{r \geq 1}\left(\inf_{\|f\|_{\mathcal{H}} \leq r} \mathbb{E}(Y - f(X))^2 + \mathbf{C}\rho(r)\right).$$

In particular, an error bound (up to log factors) follows from this result: with high probability,

$$\left\|\tilde{f} - f^*\right\|_{L_2}^2 \leq \mathbf{C}\rho(\|f^*\|_{\mathcal{H}}) = \mathbf{C}\max\left(\frac{\|f^*\|_{\mathcal{H}}^{2p/(1+p)}}{N^{1/(1+p)}}, \frac{\|f^*\|_{\mathcal{H}}^2}{N}\right). \quad (4.33)$$

One may compare this result to the one from Theorem 4.12 under assumption (4.31) even though the two procedures $\tilde{f}$ and $\hat{f}$ use different loss functions, regularization function and different search space. If assumption (4.31) holds then $\left(\sum_j \lambda_j\right)^{1/2} \leq c$ and so, one can take $r(\rho) = \left(\mathbf{C}c\rho/(\theta\sqrt{N})\right)^{1/(2\kappa)}$ and $\lambda = \mathbf{C}\sqrt{C/N}$. For such a choice of regularization parameter, Theorem 4.12 provides an error bound of the order of

$$\left\|\tilde{f} - \overline{f}\right\|_{L_2(\mu)}^2 \leq \mathbf{C}\left[\frac{\left\|\overline{f}\right\|_{\mathcal{H}_K} C}{\sqrt{N}}\right]^{1/\kappa} \quad (4.34)$$

which is almost the same as the one obtained in (4.33) when $\kappa = 1$ and p is close to 1. But our result is worse when $\kappa > 1$ and p is far from 1. This is the price that we pay by using the hinge loss – note that the quadratic loss satisfies the Bernstein condition with $\kappa = 1$ – and by fixing a regularization function which is the norm $\|\cdot\|_{\mathcal{H}_K}$ instead of fitting the regularization function in a "complexity dependent way" as in (4.32). In the last case, our procedure $\hat{f}$ does not benefit from the "real complexity" of the problem which is localized Rademacher complexities – note that we used global Rademacher complexities to fit λ and construct the complexity function $r(\cdot)$.

4.6 A review of the Bernstein and margin conditions

In order to apply the main results from Theorem 4.2 and Theorem 4.3, one has to check the Bernstein condition. This section is devoted to the study of this condition for three loss functions: the hinge loss, the quantile loss and the logistic loss. This condition has been extensively studied in Learning theory (cf. [Bartlett et al., 2003, Zhang, 2004, Mendelson, 2008, Bartlett and Mendelson, 2006, Van de Geer, 2008, Elsener and van de Geer, 2016]). We can identify mainly two approaches to study this

condition: when the class F is convex and the loss function ℓ is "strongly convex", then the risk function inherits this property and automatically satisfies the Bernstein condition (cf. [Bartlett et al., 2003]). On the other hand, for loss functions like the hinge or quantile loss, that are affine by parts, one has to use a different path. In such cases, one may go back to a statistical framework and try to check the margin assumption. As a consequence, in the latter case, the Bernstein condition is usually more restrictive and requires strong assumptions on the distribution of the observations.

4.6.1 Logistic loss

In this section, we study the Bernstein condition of the logistic loss function which is defined for every $f : \mathcal{X} \to \mathbb{R}$, $x \in \mathcal{X}$, $y \in \{-1, 1\}$ and $u \in \mathbb{R}$ by

$$\ell_f(x, y) = \tilde{\ell}(yf(x)) \text{ where } \tilde{\ell}(u) = \log(1 + \exp(-u)).$$

Function $\tilde{\ell}$ is strongly convex on every compact interval in $\mathbb{R}$. As it was first observed in [Bartlett et al., 2003, 2006], one may use this property to check the Bernstein condition for the loss function ℓ. This approach was extended to the bounded regression problem with respect to L_p loss functions ($1 < p < \infty$) in [Mendelson, 2002] and to non convex classes in [Mendelson, 2008].

In the bounded scenario, Bartlett et al. [2006] proved that the logistic loss function satisfies the Bernstein condition for $\kappa = 1$. One may therefore use that result to apply Theorem 4.3. The analysis is pretty straightforward in the bounded case. It becomes more delicate in the subgaussian scenario as considered in Theorem 4.2.

Proposition 4.1 ([Bartlett et al., 2003])**.** *Let F be a convex class of functions from $\mathcal{X}$ to $\mathbb{R}$. Assume that for every $f \in F$, $\|f\|_{L_\infty} \leq b$. Then the class F satisfies the Bernstein condition with Bernstein parameter $\kappa = 1$ and constant $A = 4\exp(2b)$.*

This result solves the problem of the Bernstein condition with respect to the logistic loss function over a convex class F of functions as long as all functions in F are uniformly bounded by some constant b. We will therefore use this result only in the bounded framework, for instance, when F is a class of linear functional indexed by a bounded set of vectors and when the design takes its values in the canonical basis.

In the subgaussian framework, one may proceed as in [Van de Geer, 2008] and assume that a statistical model holds. In that case, the Bernstein condition is reduced to the study of the Margin assumption since, in that case, the "Bayes rule" $\overline{f}$ (which is called the log-odds ratio in the case of the logistic loss function) is assumed to belong to the class F and so $f^* = \overline{f}$. The margin assumption with respect to the logistic loss function has been studied in Example 1 from [Van de Geer, 2008] but for a slightly different definition of the Margin assumption. Indeed, in [Van de Geer, 2008] only functions f in a L_∞ neighborhood of $\overline{f}$ needs to satisfy the Margin assumption whereas in Assumption 4.3 it has to be satisfied in the non-bounded set $\mathcal{C}$.

From our perspective, we do not want to make no "statistical modeling assumption". In particular, we do not want to assume that $\overline{f}$ belongs to F. We therefore have to prove the Bernstein condition when $\overline{f}$ may not belong to F. We used this result in Section 4.3 in order to obtain statistical bounds for the Logistic LASSO and Logistic Slope procedures. In those cases, F is a class of linear functionals. We now state that the Bernstein condition is satisfied for a class of linear functional when X is a standard Gaussian vector.

Proposition 4.2. *Let $F = \{\langle \cdot, t\rangle : t \in RB_{l_2}\}$ be a class of linear functionals indexed by RB_{l_2} for some radius $R \geq 1$. Let X be a standard Gaussian vector in $\mathbb{R}^d$ and let Y be a $\{-1, 1\}$ random variable. For every $f \in F$, the excess logistic risk of f, denoted by $P\mathcal{L}_f$, satisfies*

$$\mathcal{E}_{logistic}(f) = P\mathcal{L}_f \geq \frac{c_0}{R^3} \|f - f^*\|_{L_2}^2$$

where c_0 is some absolute constant.

4.6.2 Hinge loss

Unlike the logistic loss function, both the hinge loss and the quantile losses does not enjoy a strong convexity property. Therefore, one has to turn to a different approach as the one used in the previous section to check the Bernstein condition for those two loss functions.

For the hinge loss function, Bernstein condition is more stringent and is connected to the margin condition in classification. So, let us first introduce some notations specific to classification. In this setup, one is given N labeled pairs $(X_i, Y_i), i = 1, \ldots, N$ where X_i takes its values in $\mathcal{X}$ and Y_i is a label taking values in $\{-1, +1\}$. The aim is

to predict the label Y associated with X from the data when (X, Y) is distributed like the (X_i, Y_i)'s. The classic loss function considered in this setup is the $0-1$ loss function $\ell_f(x, y) = I(y \neq f(x))$ defined for any $f : \mathcal{X} \to \{-1, +1\}$. The $0-1$ loss function is not convex, this may result in some computational issues when dealing with it. A classic approach is to use a " convex relaxation function" as a surrogate to the $0-1$ loss function: note that this is a way to motivate the introduction of the hinge loss $\ell_f(x, y) = \max(1 - yf(x), 0)$. It is well known that the Bayes rules minimizes both the standard $0-1$ risk as well as the hinge risk: put $\eta(x) := \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$ and define the Bayes rule as

$$\overline{f}(x) = \operatorname{sgn}(\eta(x)), \tag{4.35}$$

then $\overline{f}$ minimizes $f \to P\ell_f$ over all measurable functions from $\mathcal{X}$ to $\mathbb{R}$ when ℓ_f is the hinge loss of f.

Let F be a class of functions from $\mathcal{X}$ to $[-1, 1]$. Assume that $\overline{f} \in F$ so that $\overline{f}$ is an oracle in F and thus (using the notations from Section 4.2) $f^* = \overline{f}$. In this situation, Margin assumption with respect to the hinge loss (cf. [Tsybakov, 2004, Lecué, 2007]) restricted to the class F and Bernstein condition (cf. Assumption 4.3) coincide. Therefore, Assumption 4.3 holds when the Margin assumption w.r.t. the hinge loss holds. According to Proposition 1 in [Lecué, 2007], the Margin assumption with respect to the hinge loss is equivalent the Margin assumption with respect to the $0-1$ loss for a class F of functions with values in $[-1, 1]$. Then, according to Proposition 1 in [Tsybakov, 2004] and [Boucheron et al., 2005] the margin assumption with respect to the $0-1$ loss with parameter κ is equivalent to

$$\begin{cases} \mathbb{P}(|\eta(X)| \leq t) \leq ct^{\frac{1}{\kappa-1}}, \forall 0 \leq t \leq 1 & \text{when } \kappa > 1 \\ |\eta(X)| \geq \tau \text{ a.s. for some } \tau > 0 & \text{when } \kappa = 1. \end{cases} \tag{4.36}$$

As a consequence, one can state the following result on the Bernstein condition for the hinge loss in the bounded case scenario.

Proposition 4.3 (Proposition 1, [Lecué, 2007])**.** *Let F be a class of functions from $\mathcal{X}$ to $[-1, 1]$. Define $\eta(x) = \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$ and assume that the Bayes rule (4.35) belongs to F. If (4.36) is satisfied for some $\kappa \geq 1$ then Assumption 4.3 holds with parameter κ for the hinge loss, and A depending on c, κ and τ (which is explicitly given in the mentioned references). In the special case when $\kappa = 1$ then $A = 1/(2\tau)$.*

Note that up to a modification of the constant A, the same result holds for functions with values in $[-b, b]$ for $b > 0$, a fact we used in Section 4.5.

4.6.3 Quantile loss

In this section, we study the Bernstein parameter of the **quantile loss** in the bounded regression model, that is when for all $f \in F, \|f\|_{L_\infty} \leq b$ a.s.. Let $\tau \in (0,1)$ and, for all $x \in \mathcal{X}$, define $\overline{f}(x)$ as the quantile of order τ of $Y|X = x$ and assume that $\bar{f}$ belongs to F, in that case, $\overline{f} = f^*$ and Bernstein condition and margin assumption are the same. Therefore one may follow the study of the margin assumption for the quantile loss in [Elsener and van de Geer, 2016] to obtain the following result.

Proposition 4.4 ([Elsener and van de Geer, 2016])**.** *Assume that for any $x \in \mathcal{X}$, it is possible to define a density f_x w.r.t the Lebesgue measure for $Y|X = x$ such that $f_x(u) \geq 1/C$ for some $C > 0$ for all $u \in \mathbb{R}$ with $|u - f^*(x)| \leq 2b$. Then the quantile loss satisfies the Bernstein's assumption with $\kappa = 1$ and $A = 2C$ over F.*

4.7 Discussion

This paper covers many aspects of the regularized empirical risk estimator (RERM) with Lipschitz loss. This property is commonly shared by many loss functions used in practice such as the hinge loss, the logistic loss or the quantile regression loss. This work offers a general method to derive estimation bounds as well as excess risk upper bounds. Two main settings are covered: the subgaussian framework and the bounded framework. The first one is illustrated by the classification problem with logistic loss. In particular, minimax rates are achieved when using the SLOPE regularization norm. The second framework is used to derive new results on matrix completion and in kernel methods.

A possible extension of this work is to study other regularization norms. In order to do that, one has to compute the complexity parameter in one of the settings and a solution of the sparsity equation. The latter usually involves to understand the sub-differential of the regularization norm and in particular its singularity points which are related to the sparsity equation.

4.8 Proof of Theorem 4.2 and Theorem 4.3

4.8.1 More general statements: Theorems 4.14 and 4.13

First, we state two theorems: Theorem 4.13 in the subgaussian setting, and Theorem 4.14 in the bounded setting. These two theorems rely on localized versions of the complexity function $r(\cdot)$ that will be defined first. Note that the localized version of $r(\cdot)$ can always be upper bounded by the simpler version used in the core of the paper. Thus, Theorem 4.2 is a direct corollary of Theorem 4.13, and Theorem 4.3 is a direct corollary of Theorem 4.14.

So let us start with a localized complexity parameters. The "statistical size" of the family of "sub-models" $(\rho B)_{\rho>0}$ is now measured by local Gaussian mean-widths in the subgaussian framework.

Definition 4.8. *Let $\theta > 0$. The **complexity parameter** is a non-decreasing function $r(\cdot)$ such that for every $\rho \geq 0$,*

$$CLw\left(\rho B \cap r(\rho) B_{L_2}\right) \leq \theta r(\rho)^{2\kappa}\sqrt{N}$$

In the boundedness case, it is written as follows.

Definition 4.9. *Let $\theta > 0$. The **complexity parameter** is a non-decreasing function $r(\cdot)$ such that for every $\rho \geq 0$,*

$$48\mathrm{Rad}\left(\rho B \cap r(\rho) B_{L_2}\right) \leq \theta r(\rho)^{2\kappa}\sqrt{N}$$

where κ is the Bernstein parameter from Assumption 4.3.

To obtain the complexity functions from Definition 4.5 and 4.7, we use the fact that $w\left(\rho B \cap r(\rho) B_{L_2}\right) \leq w(\rho B)$ and $\mathrm{Rad}\left(\rho B \cap r(\rho) B_{L_2}\right) \leq \mathrm{Rad}(\rho B)$: it indeed does not use the localization. We also set $\theta = 7/40A$ in those definitions because it is the largest value allowed in the following theorems.

Theorem 4.13. *Assume that Assumption 4.1, Assumption 4.3 and Assumption 4.4 hold where $r(\cdot)$ is a function as in Definition 4.8 for some θ such that $40A\theta \leq 7$ and assume that $\rho \to r(2\rho)/\rho$ is non-increasing. Let the regularization parameter λ be chosen such that*

$$\frac{10\theta r(2\rho)^{2\kappa}}{7\rho} < \lambda < \frac{r(2\rho)^{2\kappa}}{2A\rho}, \quad \forall \rho \geq \rho^* \tag{4.37}$$

where ρ^* satisfies (4.8)*. Then, with probability larger than*

$$1 - \sum_{j=0}^{\infty} \sum_{i \in I_j} \exp\left(- \frac{\theta^2 N (2^{(i-1)\vee 0} r(2^j \rho^*))^{4\kappa - 2}}{4C^2 L^2} \right) \tag{4.38}$$

where for all $j \in \mathbb{N}$, $I_j = \{1\} \cup \{i \in \mathbb{N}^* : 2^{i-1} r(2^j \rho^*) \leq 2^j \rho^* d_{L_2}(B)\}$, we have

$$\left\| \hat{f} - f^* \right\| \leq \rho^*, \quad \left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq r(2\rho^*)^{2\kappa}/A.$$

Proof of Theorem 4.2: Let $r(\cdot)$ be chosen as in (4.5). For this choice, one can check that the regularization parameter used for the construction of the RERM satisfies (4.37) with an adequate constant choice. Moreover, for this choice of function $r(\cdot)$ it is straightforward to lower bound the sum in the probability estimate in (4.38). The parameter λ is chosen in the middle of the range. ■

The bounded case is in the same spirit.

Theorem 4.14. *Assume that Assumption 4.1, Assumption 4.3 and Assumption 4.5 hold where $r(\cdot)$ is a function as in Definition 4.9 for some θ such that $40A\theta \leq 7$ and assume that $\rho \to r(2\rho)/\rho$ is non-increasing. Let the regularization parameter λ be chosen such that*

$$\frac{10\theta r(2\rho)^{2\kappa}}{7\rho} < \lambda < \frac{r(2\rho)^{2\kappa}}{2A\rho}, \quad \forall \rho \geq \rho^* \tag{4.39}$$

where ρ^* satisfies (4.8)*. Then, with probability larger than*

$$1 - 2\sum_{j=0}^{\infty} \sum_{i \in I_j} \exp\left(-c_0 \theta^2 N (2^i r(2^{j+1} \rho^*))^{4\kappa - 2} \right) \tag{4.40}$$

where $c_0 = 1/\max\left(48, 207\theta b^{2\kappa-1}\right)$ and for all $j \in \mathbb{N}$, $I_j := \{1\} \cup \{i \in \mathbb{N}^* : 2^{i-1} r(2^j \rho^*) \leq \min(2^j \rho^* d_{L_2}(B), b)\}$, we have

$$\left\| \hat{f} - f^* \right\| \leq \rho^*, \quad \left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq r(2\rho^*)^{2\kappa}/A.$$

The proof of Theorem 4.3 is identical to the one of Theorem 4.2 and we do not reproduce it here.

4.8.2 Proofs of Theorems 4.14 and 4.13

Proof of Theorem 4.13 and and Theorem 4.14 follow the same strategy. They are split into two parts. First, we identify an event onto which the statistical behavior of the regularized estimator $\hat{f}$ can be controlled using only deterministic arguments. Then, we prove that this event holds with a probability at least as large as the one in (4.38) in the case of Theorem 4.13 and as in (4.40) in the case of Theorem 4.14. We first introduce this event which is common to the subgaussian and the bounded setups:

$$\Omega_0 := \left\{ \left| (P - P_N)\mathcal{L}_f \right| \leq \theta \max\left(r(2\max(\|f - f^*\|, \rho^*))^{2\kappa}, \|f - f^*\|_{L_2}^{2\kappa} \right) : \text{ for all } f \in F \right\}$$

where θ is a parameter appearing in the definition of $r(\cdot)$ in Definition 4.8 and Definition 4.9, $\kappa \geq 1$ is the Bernstein parameter from Definition 4.3 and ρ^* is a radius satisfying the sparsity Equation (4.8).

Proposition 4.5. *Let λ be as in* (4.37) *(or equivalently as in* (4.39)*) and let ρ^* satisfy* (4.8)*, on the event Ω_0, one has*

$$\left\| \hat{f} - f^* \right\| \leq \rho^*, \quad \left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\rho^*) \text{ and } \mathcal{E}(\hat{f}) \leq \theta r(2\rho^*)^{2\kappa}.$$

Proof. Denote $\hat{\rho} = \left\| \hat{f} - f^* \right\|$. We first prove that $\hat{\rho} < \rho^*$. To that end, we assume that the reverse inequality holds and show some contradiction. Assume that $\hat{\rho} \geq \rho^*$. Since $\rho \to r(2\rho)/\rho$ is non-increasing then by Lemma 4.4, $\rho \to \Delta(\rho)/\rho$ is non-decreasing and so we have

$$\frac{\Delta(\hat{\rho})}{\hat{\rho}} \geq \frac{\Delta(\rho^*)}{\rho^*} \geq \frac{4}{5}.$$

Now, we consider two cases: either $\left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\hat{\rho})$ or $\left\| \hat{f} - f^* \right\|_{L_2} > r(2\hat{\rho})$.

First assume that $\left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\hat{\rho})$. Since $\Delta(\hat{\rho}) \geq 4\hat{\rho}/5$ and $h = \hat{f} - f^* \in \hat{\rho}S \cap r(2\hat{\rho})B_{L_2}$, it follows from the definition of the sparsity parameter $\Delta(\hat{\rho})$ that there exists some $f \in F$ such that $\|f - f^*\| \leq \hat{\rho}/20$ and for which

$$\|f + h\| - \|f\| \geq \frac{4\hat{\rho}}{5}.$$

It follows that

$$\left\| \hat{f} \right\| - \|f^*\| = \|f^* + h\| - \|f^*\| \geq \|f + h\| - \|f\| - 2\|f - f^*\| \geq \frac{4\hat{\rho}}{5} - \frac{\hat{\rho}}{10} = \frac{7\hat{\rho}}{10}.$$

Let us now introduce the excess regularized loss: for all $f \in F$,

$$\mathcal{L}_f^\lambda = \mathcal{L}_f + \lambda(\|f\| - \|f^*\|) = (\ell_f + \lambda \|f\|) - (\ell_{f^*} + \lambda \|f^*\|).$$

On the event Ω_0, we have

$$P_N \mathcal{L}_{\hat{f}}^\lambda = P_N \mathcal{L}_{\hat{f}} + \lambda \left(\left\| \hat{f} \right\| - \|f^*\| \right) \geq (P_N - P)\mathcal{L}_{\hat{f}} + \lambda \left(\left\| \hat{f} \right\| - \|f^*\| \right)$$
$$\geq -\theta \max \left(r(2\hat{\rho})^{2\kappa}, \left\| \hat{f} - f^* \right\|_{L_2}^{2\kappa} \right) + \frac{7\lambda\hat{\rho}}{10} = -\theta r(2\hat{\rho})^{2\kappa} + \frac{7\lambda\hat{\rho}}{10} > 0$$

because by definition of λ, $7\lambda\hat{\rho} > 10\theta r(2\hat{\rho})^{2\kappa}$. Therefore, $P_N \mathcal{L}_{\hat{f}}^\lambda > 0$. But, by construction, one has $P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0$.

Then, assume that $\left\| \hat{f} - f^* \right\|_{L_2} > r(2\hat{\rho})$. In particular, $f \in \mathcal{C}$ where $\mathcal{C}$ is the set introduced in 4.7 below Assumption 4.3. By definition of $\hat{f}$ we have $P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0$ so it follows from Assumption 4.3 that

$$\left\| \hat{f} - f^* \right\|_{L_2}^{2\kappa} \leq AP\mathcal{L}_{\hat{f}} = A \left[(P - P_N)\mathcal{L}_{\hat{f}} + P_N \mathcal{L}_{\hat{f}}^\lambda + \lambda \left(\|f^*\| - \left\| \hat{f} \right\| \right) \right]$$
$$\leq A\theta \max \left(r(2\hat{\rho})^{2\kappa}, \left\| \hat{f} - f^* \right\|_{L_2}^{2\kappa} \right) + A\lambda \left\| \hat{f} - f^* \right\| = A\theta \left\| \hat{f} - f^* \right\|_{L_2}^{2\kappa} + A\lambda\hat{\rho}. \tag{4.41}$$

Hence, if $A\theta \leq 1/2$ then

$$r(2\hat{\rho})^{2\kappa} \leq \left\| \hat{f} - f^* \right\|_{L_2}^{2\kappa} \leq 2A\lambda\hat{\rho}.$$

But, by definition of λ one has $r(2\hat{\rho})^{2\kappa} > 2A\lambda\hat{\rho}$.

Therefore, none of the two cases is possible when one assumes that $\hat{\rho} \geq \rho^*$ and so we necessarily have $\hat{\rho} < \rho^*$.

Now, assuming that $\left\| \hat{f} - f^* \right\|_{L_2} > r(2\rho^*)$ and following (4.41) step by step also leads to a contradiction, so $\left\| \hat{f} - f^* \right\|_{L_2} \leq r(2\rho^*)$.

Next, we prove the result for the excess risk. One has

$$P_N \mathcal{L}_{\hat{f}}^\lambda = P_N \mathcal{L}_{\hat{f}} + \lambda \left(\left\| \hat{f} \right\| - \|f^*\| \right) = (P_N - P)\mathcal{L}_{\hat{f}} + P\mathcal{L}_{\hat{f}} + \lambda \left(\left\| \hat{f} \right\| - \|f^*\| \right)$$
$$\geq -\theta \max \left(r(2\rho^*)^{2\kappa}, \left\| \hat{f} - f^* \right\|_{L_2}^{2\kappa} \right) + P\mathcal{L}_{\hat{f}} - \lambda\hat{\rho} \geq -\theta r(2\rho^*)^{2\kappa} - \lambda\rho^* + P\mathcal{L}_{\hat{f}}$$

$$\geq -\left(\theta + \frac{1}{2A}\right) r(2\rho^*)^{2\kappa} + P\mathcal{L}_{\hat{f}} \geq \frac{-r(2\rho^*)^{2\kappa}}{A} + P\mathcal{L}_{\hat{f}}.$$

In particular, if $P\mathcal{L}_{\hat{f}} > r(2\rho^*)^{2\kappa}/A$ then $P_N\mathcal{L}_{\hat{f}}^{\lambda} > 0$ which is not possible by construction of $\hat{f}$ so we necessarily have $P\mathcal{L}_{\hat{f}} \leq r(2\rho^*)^{2\kappa}/A$. ∎

Proposition 4.5 shows that $\hat{f}$ satisfies some estimation and prediction properties on the event Ω_0. Next, we prove that Ω_0 holds with large probability in both subgaussian and bounded frameworks. We start with the subgaussian framework. To that end, we introduce several tools.

Recall that the ψ_2-norm of a real valued random variable Z is defined by

$$\|Z\|_{\psi_2} = \inf\left\{c > 0 : \mathbb{E}\psi_2(|Z|/c) \leq \psi_2(1)\right\}$$

where $\psi_2(u) = \exp(u^2) - 1$ for all $u \geq 0$. The space L_{ψ_2} of all real valued random variables with finite ψ_2-norm is called the Orlicz space of subgaussian variables. We refer the reader to [Rao and Ren, 1991, 2002] for more details on Orlicz spaces.

We recall several facts on the ψ_2-norm and subgaussian processes. First, it follows from Theorem 1.1.5 from [Chafaï et al., 2012] that $\|Z\|_{\psi_2} \leq \max(K_0, K_1)$ if

$$\mathbb{E}\exp(\lambda|Z|) \leq \exp\left(\lambda^2 K_1^2\right), \quad \forall \lambda \geq 1/K_0. \tag{4.42}$$

It follows from Lemma 1.2.2 from [Chafaï et al., 2012] that, if Z is a centered ψ_2 random variable then, for all $\lambda > 0$,

$$\mathbb{E}\exp\left(\lambda Z\right) \leq \exp\left(e\lambda^2 \|Z\|_{\psi_2}^2\right). \tag{4.43}$$

Then, it follows from Theorem 1.2.1 from [Chafaï et al., 2012] that if $Z_1, \ldots, Z_N$ are independent centered real valued random variables then

$$\left\|\sum_{i=1}^N Z_i\right\|_{\psi_2} \leq 16\left(\sum_{i=1}^N \|Z_i\|_{\psi_2}^2\right)^{1/2}. \tag{4.44}$$

Finally, let us turn to some properties of subgaussian processes. Let (T, d) be a pseudo-metric space. Let $(X_t)_{t\in T}$ be a random process in L_{ψ_2} such that for all $s, t \in T$, $\|X_t - X_s\|_{\psi_2} \leq d(s, t)$. It follows from the comment below Theorem 11.2 p.300 in [Ledoux and Talagrand, 1991] that for all measurable set A and all $s, t \in T$,

$$\int_A |X_s - X_t| d\mathbb{P} \leq d(s, t)\mathbb{P}(A)\psi_2^{-1}\left(\frac{1}{\mathbb{P}(A)}\right).$$

Therefore, it follows from equation (11.14) in [Ledoux and Talagrand, 1991] that for every $u > 0$,

$$\mathbb{P}\left(\sup_{s,t\in T}|X_s - X_t| > c_0(\gamma_2 + Du)\right) \leq \psi_2(u)^{-1} \tag{4.45}$$

where D is the diameter of (T, d), c_0 is an absolute constant and γ_2 is the majorizing measure integral $\gamma(T, d; \psi_2)$ (cf. Chapter 11 in [Ledoux and Talagrand, 1991]). When T is a subset of L_2 and d is the natural metric of L_2 it follows from the majorizing measure theorem that $\gamma_2 \leq c_1 w(T)$ (cf. Chapter 1 in [Talagrand, 2005]).

Lemma 4.3. *Assume that Assumption 4.1 and Assumption 4.4 hold. Let $F' \subset F$ then for every $u > 0$, with probability at least $1 - 2\exp(-u^2)$*

$$\sup_{f,g\in F'}|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{c_0 L}{\sqrt{N}}\left(w(F') + u d_{L_2}(F')\right)$$

where d is the L_2 metric and $d_{L_2}(F')$ is the diameter of (F', d).

Proof. To prove Lemma 4.3, it is enough to show that $((P - P_N)\mathcal{L}_f)_{f\in F'}$ has $(L/\sqrt{N})$-subgaussian increments and then to apply (4.45) where $\gamma_2 \sim w(F')$ in this case.

Let us prove that for some absolute constant c_0: for all $f, g \in F'$,

$$\|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\|_{\psi_2} \leq c_0(L/\sqrt{N})\,\|f - g\|_{L_2}$$

It follows from (4.44) that

$$\left\|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\right\|_{\psi_2} \leq 16\left(\sum_{i=1}^{N}\frac{\left\|(\mathcal{L}_f - \mathcal{L}_g)(X_i, Y_i) - \mathbb{E}(\mathcal{L}_f - \mathcal{L}_g)\right\|_{\psi_2}^2}{N^2}\right)^{1/2} = \frac{16}{\sqrt{N}}\left\|\zeta_{f,g}\right\|_{\psi_2}.$$

where $\zeta_{f,g} = (\mathcal{L}_f - \mathcal{L}_g)(X, Y) - \mathbb{E}(\mathcal{L}_f - \mathcal{L}_g)$.Therefore, it only remains to show that $\|\zeta_{f,g}\|_{\psi_2} \leq c_1 L\,\|f - g\|_{L_2}$.

It follows from (4.42), that the last inequality holds if one proves that for all $\lambda \geq c_1/(L\,\|f - g\|_{L_2})$,

$$\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right) \leq \exp(c_2\lambda^2 L^2\,\|f - g\|_{L_2}^2) \tag{4.46}$$

for some absolute constants c_1 and c_2. To that end, it is enough to prove that, for some absolute constant c_3 – depending only on c_1 and c_2 – and all $\lambda > 0$,

$$\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right) \leq 2\exp(c_3\lambda^2 L^2\,\|f - g\|_{L_2}^2).$$

Note that if Z is a real valued random variable and ϵ is a Rademacher variable independent of Z then $\mathbb{E}\exp(|Z|) \leq 2\exp(\epsilon Z)$. Hence, it follows from a symmetrization argument (cf. Lemma 6.3 in [Ledoux and Talagrand, 1991]), (a simple version of) the contraction principle (cf. Theorem 4.4 in [Ledoux and Talagrand, 1991]) and (4.43) that, for all $\lambda > 0$,

$$\mathbb{E}\exp\left(\lambda|\zeta_{f,g}|\right) \leq 2\mathbb{E}\exp(\lambda\epsilon\zeta_{f,g}) \leq 2\mathbb{E}\exp\left(2\lambda\epsilon(\mathcal{L}_f - \mathcal{L}_g)(X,Y)\right)$$
$$\leq 2\mathbb{E}\exp\left(2\lambda\epsilon(f-g)(X)\right) \leq 2\mathbb{E}\exp\left(c_4\lambda^2 L^2 \|f-g\|_{\psi_2}^2\right)$$

where ϵ is a Rademacher variable independent of (X,Y) and where we used in the last but one inequality that $|\mathcal{L}_f(X,Y) - \mathcal{L}_g(X,Y)| \leq |f(X) - g(X)|$ a.s.. ■

Proposition 4.6. *We assume that Assumption 4.1, 4.4 and 4.3 hold. Then the probability measure of Ω_0 is at least as large as the one in* (4.38).

Proof. The proof is based on a peeling argument (cf. [van de Geer, 2000]) with respect to the two distances naturally associated with this problem: the regularization norm $\|\cdot\|$ and the L_2-norm $\|\cdot\|_{L_2}$ associated with the design X. The peeling according to $\|\cdot\|$ is performed along the radii $\rho_j = 2^j\rho^*$ for $j \in \mathbb{N}$ and the peeling according to $\|\cdot\|_{L_2}$ is performed within the class $\{f \in F : \|f - f^*\| \leq \rho_j\} := f^* + \rho_j B$ along the radii $2^i r(\rho_j)$ for all $i = 0, 1, 2, \cdots$ up to a radius such that $2^i r(\rho_j)$ becomes larger than the radius of $f^* + \rho_j B$ in L_2, that is for all $i \in I_j$.

We introduce the following partition of the class F. We first introduce the "true model", i.e. the subset of F where we want to show that $\hat{f}$ belongs to with high probability:

$$F_{0,0} = \left\{f \in F : \|f - f^*\| \leq \rho_0 \text{ and } \|f - f^*\|_{L_2} \leq r(\rho_0)\right\}$$

(note that $\rho_0 = \rho^*$). Then we peel the remaining set $F \backslash F_{0,0}$ according to the two norms: for every $i \in I_0$,

$$F_{0,i} = \left\{f \in F : \|f - f^*\| \leq \rho_0 \text{ and } 2^{i-1}r(\rho_0) < \|f - f^*\|_{L_2} \leq 2^i r(\rho_0)\right\},$$

for all $j \geq 1$,

$$F_{j,0} = \left\{f \in F : \rho_{j-1} < \|f - f^*\| \leq \rho_j \text{ and } \|f - f^*\|_{L_2} \leq r(\rho_j)\right\}$$

and for every integer $i \in I_j$,

$$F_{j,i} = \left\{ f \in F : \rho_{j-1} < \|f - f^*\| \leq \rho_j \text{ and } 2^{i-1} r(\rho_j) < \|f - f^*\|_{L_2} \leq 2^i r(\rho_j) \right\}.$$

We also consider the sets $F^*_{j,i} = \rho_j B \cap (2^i r(\rho_j)) B_{L_2}$ for all integers i and j.

Let j and $i \in I_j$ be two integers. It follows from Lemma 4.3 that for any $u > 0$, with probability larger than $1 - 2\exp(-u^2)$,

$$\sup_{f \in F_{j,i}} |(P - P_N)\mathcal{L}_f| \leq \sup_{f,g \in F^*_{j,i} + f^*} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{c_0 L}{\sqrt{N}} \left(w(F^*_{j,i}) + u d_{L_2}(F^*_{j,i}) \right) \tag{4.47}$$

where $d_{L_2}(F^*_{j,i}) \leq 2^{i+1} r(\rho_j)$.

Note that for any $\rho > 0$, $h : r \to w(\rho B \cap r B_{L_2})/r$ is non-increasing (cf. Lemma 4.5 in the Appendix) and note that, by definition of $r(\rho)$ (cf. Definition 4.8), $h(r(\rho)) \leq \theta r(\rho)^{2\kappa - 1} \sqrt{N}/(CL)$. Since $h(\cdot)$ is non-increasing, we have $w(F^*_{j,i})/(2^i r(\rho_j)) \leq h(2^i r(\rho_j)) \leq h(r(\rho_j)) \leq \theta r(\rho_j)^{2\kappa - 1} \sqrt{N}/(CL)$ and so $w(F^*_{j,i}) \leq \theta 2^i r(\rho_j)^{2\kappa} \sqrt{N}/(CL)$. Therefore, it follows from (4.47) for $u = \theta \sqrt{N} (2^{(i-1)\vee 0} r(\rho_j))^{2\kappa - 1}/(2CL)$, if $C \geq 4c_0$ then, with probability at least

$$1 - 2\exp\left(-\theta^2 N (2^{(i-1)\vee 0} r(\rho_j))^{4\kappa - 2}/(4C^2 L^2) \right), \tag{4.48}$$

for every $f \in F_{j,i}$,

$$|(P - P_N)\mathcal{L}_f| \leq \theta (2^{(i-1)\vee 0} r(\rho_j))^{2\kappa} \leq \theta \max\left(r(2\max(\|f - f^*\|, \rho^*))^{2\kappa}, \|f - f^*\|_{L_2}^{2\kappa} \right).$$

The result follows from a union bound. ∎

Now we turn to the proof of Theorem 4.13 under the boundedness assumption. The proof follows the same strategy as in the "subgaussian case": we first use Proposition 4.5 and then show (under the boundedness assumption) that event Ω_0 holds with probability at least as large as the one in (4.40).

Similar to Proposition 4.6, we prove the following result under the boundedness assumption.

Proposition 4.7. *We assume that Assumption 4.1, 4.5 and 4.3 hold. Then the probability measure of Ω_0 is at least as large as the one in* (4.40).

Proof. Using the same notation as in the proof of Proposition 4.6, we have for any integer j and i such that $2^i r(\rho_j) \leq b$ that by Talagrand's concentration inequality: for any $x > 0$, with probability larger than $1 - 2e^{-x}$,

$$Z_{j,i} \leq 2\mathbb{E}Z_{j,i} + \sigma(\mathcal{L}_{F_{j,i}})\sqrt{\frac{8x}{N}} + \frac{69\left\|\mathcal{L}_{F_{j,i}}\right\|_\infty x}{2N} \tag{4.49}$$

where

$$Z_{j,i} = \sup_{f\in F_{j,i}} |(P-P_N)\mathcal{L}_f|, \quad \sigma(\mathcal{L}_{F_{j,i}}) = \sup_{f\in F_{j,i}} \sqrt{\mathbb{E}\mathcal{L}_f^2} \text{ and } \left\|\mathcal{L}_{F_{j,i}}\right\|_\infty = \sup_{f\in F_{j,i}} \|\mathcal{L}_f\|_\infty .$$

By the Lipschitz assumption, one has

$$\sigma(\mathcal{L}_{F_{j,i}}) \leq 2^{i+1} r(\rho_j) \text{ and } \left\|\mathcal{L}_{F_{j,i}}\right\|_\infty \leq 2b.$$

Therefore, it only remains to upper bound the expectation $\mathbb{E}Z_{j,i}$. Let $\epsilon_1, \ldots, \epsilon_N$ be a N i.i.d. Rademacher variables independent of the (X_i, Y_i)'s. For all function f, we set

$$P_{N,\epsilon} f = \frac{1}{N}\sum_{i=1}^N \epsilon_i f(X_i)$$

It follows from a symmetrization and a contraction argument (cf. Chapter 4 in [Ledoux and Talagrand, 1991]) that

$$\mathbb{E}Z_{j,i} \leq 4\mathbb{E}\sup_{f\in F_{j,i}} |P_{N,\epsilon}(f-f^*)| \leq \frac{4\mathrm{Rad}(\rho_j B \cap (2^i r(\rho_j))B_{L_2})}{\sqrt{N}} \leq (\theta/12)2^i r(\rho_j)^{2\kappa}.$$

Now, we take $x = c_2\theta^2 N(2^{i-1}r(\rho_j))^{4\kappa-2}$ in (4.49) and note that $2^i r(\rho_j) \leq b$ and $\kappa \geq 1$: with probability larger than

$$1 - 2\exp(-c_2\theta N(2^i r(\rho_j))^{4\kappa-2}), \tag{4.50}$$

for any $f \in F_{j,i}$,

$$\begin{aligned}
|(P-P_N)\mathcal{L}_f| &\leq \theta 2^{i-1} r(\rho_j)^{2\kappa}/3 + 2\sqrt{8c_2}\theta\left(2^{i-1}r(\rho_j)\right)^{2\kappa} + 69c_2\theta^2 b(2^{i-1}r(\rho_j))^{4\kappa-2} \\
&\leq \theta\left(2^{(i-1)\vee 0} r(\rho_j)\right)^{2\kappa}\left[\frac{1}{3} + 2\sqrt{8c_2} + 69c_2\theta b(2^i r(\rho_j))^{2\kappa-2}\right] \\
&\leq \theta\left(2^{(i-1)\vee 0} r(\rho_j)\right)^{2\kappa}\left[\frac{1}{3} + 2\sqrt{8c_2} + 69c_2\theta b^{2\kappa-1}\right]
\end{aligned}$$

$$\leq \theta \max\left(r(2\max(\|f-f^*\|,\rho^*))^{2\kappa}, \|f-f^*\|_{L_2}^{2\kappa}\right)$$

if c_2 is defined by

$$c_2 = \min\left(\frac{1}{48}, \frac{1}{207\theta b^{2\kappa-1}}\right). \tag{4.51}$$

We conclude with a union bound. ■

4.9 Proof of Theorem 4.8

For the sake of simplicity, assume that $m \geq T$ so $\max(m,T) = m$. Fix $r \in \{1,\dots,T\}$. Fix $x > 0$ such that $\exp(x)/[1+\exp(x)] \leq b$, we define the set of matrices

$$\mathcal{C}_x = \left\{A \in \mathbb{R}^{m\times r} : \forall (p,q), A_{p,q} \in \{0,x\}\right\}$$

and

$$\mathcal{M}_x = \{A \in \mathbb{R} : A = (B|\dots|B|O), B \in \mathcal{C}_x\}$$

where the block B is repeated $\lfloor T/r \rfloor$ times (this construction is taken from [Koltchinskii et al., 2011]). Varshamov-Gilbert bound (Lemma 2.9 in [Tsybakov, 2009]) implies that there is a finite subset $\mathcal{M}_x^0 \subset \mathcal{M}_x$ with $\mathrm{card}(\mathcal{M}_x^0) \geq 2^{rm/8}+1$ with $0 \in \mathcal{M}_x^0$, and for any distinct $A, B \in \mathcal{M}_x^0$,

$$\|A-B\|_{S_2}^2 \geq \frac{mr\lfloor T/r \rfloor}{8} x^2 \geq \frac{mT}{16} x^2$$

and so

$$\frac{1}{mT}\|A-B\|_{S_2}^2 \geq \frac{x^2}{16}.$$

Then, for $A \in \mathcal{M}_x^0 \setminus \{0\}$,

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_0, \mathbb{P}_A) &= \frac{n}{mT}\sum_{i=1}^{m}\sum_{j=1}^{T}\left[\frac{1}{2}\log\left(\frac{1+\exp(M_{i,j})}{2\exp(M_{i,j})}\right) + \frac{1}{2}\log\left(\frac{1+\exp(M_{i,j})}{2}\right)\right] \\
&= \frac{n}{mT}\sum_{i=1}^{m}\sum_{j=1}^{T}\left[\log\left(\frac{1+\exp(M_{i,j})}{2}\right) - \frac{1}{2}M_{i,j}\right] \\
&\leq n\left[\log\left(\frac{1+\exp(x)}{2}\right) - \frac{1}{2}x\right] \\
&\leq c(b) n x^2
\end{aligned}$$

where $c(b) > 0$ is a constant that depends only on b. So:

$$\frac{1}{\mathrm{card}(\mathcal{M}_x^0) - 1} \sum_{A \in \mathcal{M}_x^0} \mathcal{K}(\mathbb{P}_0, \mathbb{P}_A) \leq c(b) n x^2 \leq c(b) \log(\mathrm{card}(\mathcal{M}_x^0) - 1)$$

as soon as we choose

$$x \leq \sqrt{\frac{\log(\mathrm{card}(\mathcal{M}_x^0) - 1)}{n}} \leq \sqrt{\frac{rm \log(2)}{8n}}$$

(note that the condition $n \geq rm \log(2)/(8b^2)$ implies that $\exp(x)/[1 + \exp(x)] \leq b$). Then, Theorem 2.5 in [Tsybakov, 2009] leads to the existence of $\beta, c > 0$ such that

$$\inf_{\widehat{M}} \sup_{A \in \mathcal{M}_x^0} \mathbb{P}_A \left(\frac{1}{mT} \|\widehat{M} - A\|_{S_2}^2 \geq c \frac{mr}{N} \right) \geq \beta.$$

■

4.10 Proof of Theorem 4.10

For the sake of simplicity, assume that $m \geq T$ so $\max(m, T) = m$. Fix $r \in \{2, \ldots, T\}$ and assume that $rT \leq N \leq mT$.

We recall that $\{E_{p,q} : 1 \leq p \leq m, 1 \leq p \leq T\}$ is the canonical basis of $\mathbb{R}^{m \times T}$. We consider the following "blocks of coordinates": for every $1 \leq k \leq r - 1$ and $1 \leq l \leq T$,

$$B_{kl} = \left\{ E_{p,l} : \frac{(k-1)mT}{N} + 1 \leq p < \frac{kmT}{N} + 1 \right\}$$

(note that $(r-1)mT/N + 1 \leq m$ when $rT \leq N \leq mT$). We also introduce the "blocks" of "remaining" coordinates:

$$B_0 = \left\{ E_{p,q} : \frac{(r-1)mT}{N} + 1 \leq p, 1 \leq q \leq T \right\}$$

For every $\sigma = (\sigma_{kl}) \in \{0, 1\}^{(r-1) \times T}$, we denote by $\mathbb{P}_\sigma$ the probability distribution of a pair (X, Y) taking its values in $\mathbb{R}^{m \times T} \times \{-1, 1\}$ where X is uniformly distributed over the basis $\{E_{p,q} : 1 \leq p \leq m, 1 \leq p \leq T\}$ and for every $(p, q) \in \{1, \ldots, m\} \times \{1, \ldots, T\}$,

$$\mathbb{P}_\sigma[Y = 1 | X = E_{p,q}] = \begin{cases} \sigma_{kl} & \text{if } E_{p,q} \in B_{kl} \\ 1 & \text{otherwise.} \end{cases}$$

We also introduce $\eta_\sigma(E_{p,q}) = \mathbb{E}[Y = 1|X = E_{p,q}] = 2\mathbb{P}_\sigma[Y = 1|X = E_{p,q}] - 1$. It follows from [Zhang, 2004] that the Bayes rules minimizes the Hinge risk, that is $f^*_\sigma \in \arg\min_f \mathbb{E}_\sigma(Y - f(X))_+$, where the minimum runs over all measurable functions and $\mathbb{E}_\sigma$ denotes the expectation w.r.t. (X, Y) when $(X, Y) \sim \mathbb{P}_\sigma$, is achieved by $f^*_\sigma = \mathrm{sgn}(\eta_\sigma(\cdot))$. Therefore, $f^*_\sigma(\cdot) = \langle M^*_\sigma, \cdot\rangle$ where for every $(p, q) \in \{1, \dots, m\} \times \{1, \dots, T\}$,

$$(M^*_\sigma)_{pq} = \begin{cases} 2\sigma_{kl} - 1 & \text{if } E_{p,q} \in B_{kl} \\ 1 & \text{otherwise.} \end{cases} = \eta_\sigma(E_{p,q}).$$

In particular, M^*_σ has a rank at most equal to r.

Let $\sigma = (\sigma_{p,q}), \sigma' = (\sigma'_{pq})$ be in $\{0, 1\}^{(r-1)T}$. We denote by $\rho(\sigma, \sigma')$ the Hamming distance between σ and σ' (i.e. the number of times the coordinates of σ and σ' are different). We denote by $H(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'})$ the Hellinger distance between the probability measures $\mathbb{P}_\sigma$ and $\mathbb{P}_{\sigma'}$. We have

$$H(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'}) = \int \left(\sqrt{d\mathbb{P}_\sigma} - \sqrt{d\mathbb{P}_{\sigma'}}\right)^2 = \frac{2\rho(\sigma, \sigma')}{N}.$$

Then, if $\rho(\sigma, \sigma') = 1$, it follows that (cf. Section 2.4 in [Tsybakov, 2009]),

$$H^2(\mathbb{P}^{\otimes N}_\sigma, \mathbb{P}^{\otimes N}_{\sigma'}) = 2\left(1 - \left(1 - \frac{H^2(\mathbb{P}_\sigma, \mathbb{P}_{\sigma'})}{2}\right)^N\right) = 2\left(1 - \left(1 - \frac{1}{N}\right)^N\right) \le 2(1-e^{-2}) := \alpha.$$

Now, it follows from Theorem 2.12 in [Tsybakov, 2009], that

$$\inf_{\hat\sigma} \max_{\sigma \in \{0,1\}^{(r-1)T}} \mathbb{E}^{\otimes N}_\sigma \|\hat\sigma - \sigma\|_{l_1} \ge \frac{(r-1)T}{8}\left(1 - \sqrt{\alpha(1 - \alpha/4)}\right) \quad (4.52)$$

where the infimum $\inf_{\hat\sigma}$ runs over all measurable functions $\hat\sigma$ of the data $(X_i, Y_i)_{i=1}^N$ with values in $\mathbb{R}$ (note that Theorem 2.12 in [Tsybakov, 2009] is stated for functions $\hat\sigma$ taking values in $\{0, 1\}^{(r-1)T}$ but its is straightforward to extend this result to any $\hat\sigma$ valued in $\mathbb{R}$) and $\mathbb{E}^{\otimes N}_\sigma$ denotes the expectation w.r.t. those data distributed according to $\mathbb{P}^{\otimes N}_\sigma$.

Now, we lower bound the excess risk of any estimator. Let $\hat f$ be an estimator with values in $\mathbb{R}$. Using a truncation argument it is not hard to see that one can restrict the values of $\hat f$ to $[-1, 1]$. In that case, We have

$$\mathcal{E}_{hinge}(\hat f) = \mathbb{E}\left[|2\eta_\sigma(X) - 1||\hat f(X) - f^*_\sigma(X)|\right] = \mathbb{E}|\hat f(X) - f^*_\sigma(X)|$$
$$= \sum_{p,q} |\hat f(E_{p,q}) - f^*_\sigma(E_{p,q})|\mathbb{P}[X = E_{p,q}]$$

$$\geq \sum_{kl} \frac{1}{mT} \sum_{E_{p,q} \in B_{kl}} |\hat{f}(E_{p,q}) - (2\sigma_{pq} - 1)| \geq \frac{2}{N} \sum_{kl} |\hat{\sigma}_{kl} - \sigma_{pq}|$$

where $\hat{\sigma}_{kl}$ is the mean of $\{(\hat{f}(E_{p,q}) + 1)/2 : E_{p,q} \in B_{kl}\}$. Then we obtain,

$$\inf_{\hat{f}} \sup_{\sigma \in \{0,1\}^{(r-1)T}} \mathbb{E}_{\sigma}^{\otimes N} \mathcal{E}_{hinge}(\hat{f}) \geq \frac{2}{N} \inf_{\hat{\sigma}} \max_{\sigma \in \{0,1\}^{(r-1)T}} \mathbb{E}_{\sigma}^{\otimes N} \|\hat{\sigma} - \sigma\|_{l_1}$$

and, using (4.52), we get

$$\inf_{\hat{f}} \sup_{\sigma \in \{0,1\}^{(r-1)T}} \mathbb{E}_{\sigma}^{\otimes N} \mathcal{E}_{hinge}(\hat{f}) \geq c_0 \frac{rT}{N}$$

for $c_0 = \left(1 - \sqrt{\alpha(1 - \alpha/4)}\right)/4$.

■

4.11 Proofs of Section 4.6

4.11.1 Proof of Section 4.6.1

The proof of Proposition 4.1 may be found in several papers (cf., for instance, [Bartlett et al., 2003]). Let us recall this argument since we will be using it at a starting point to prove the Bernstein condition in the subgaussian case.

Proof of Proposition 4.1: The logistic risk of a function $f : \mathcal{X} \to \mathbb{R}$ can be written as $P\ell_f = \mathbb{E}[g(X, f(X))]$ where for all $x, a \in \mathbb{R}$, $g(x, a) := ((1 + \eta(x))/2) \log(1 + e^{-a}) + ((1 - \eta(x))/2) \log(1 + e^{a})$ and $\eta(x) = \mathbb{E}[Y|X = x]$ is the conditional expectation of Y given $X = x$.

Since f^* minimizes $f \to P\ell_f$ over the convex class F, one has by the first order condition that for every $f \in F$, $\mathbb{E}\partial_2 g(X, f^*(X))(f - f^*)(X) \geq 0$. Therefore, it follows from a second order Taylor expansion that the excess logistic loss of every $f \in F$ is such that

$$\mathcal{E}_{logistic}(f) = P\mathcal{L}_f \geq \mathbb{E}\left[(f(X) - f^*(X))^2 \int_0^1 (1 - u)\delta(f^*(X) + u(f - f^*)(X))du\right] \tag{4.53}$$

where $\delta(u) = \partial_2^2 g(x, u) = e^u/(1 + e^u)^2$ for every $u \in \mathbb{R}$.

Since $|f^*(X)|, |f(X)| \leq b$ a.s. then for every $u \in [0, 1]$, $|f^*(X) + u(f - f^*)(X)| \leq 2b$, a.s. and since $\delta(v) \geq \delta(2b) \geq \exp(-2b)/4$ for every $|v| \leq 2b$, it follows from (4.53) that $P\mathcal{L}_f \geq \delta(2b) \|f - f^*\|_{L_2}^2$. ■

Proof of Proposition 4.2: Let $t^* \in RB_{l_2}$ be such that $f^* = \langle \cdot, t^* \rangle$, where f^* is an oracle in $F = \{\langle \cdot, t \rangle : t \in RB_{l_2}\}$ w.r.t. the logistic loss risk. Let $f = \langle \cdot, t \rangle \in F$ for some $t \in RB_{l_2}$. It follows from (4.53) that the excess logistic risk of f satisfies

$$P\mathcal{L}_f \geq \int_0^1 \mathbb{E}\left[\langle X, t^* - t\rangle^2 \, \delta\left(\langle X, t^* + u(t - t^*)\rangle\right)\right] du.$$

The result will follow if one proves that for every $t_0, t \in \mathbb{R}^d$,

$$\mathbb{E}\left[\langle X, t\rangle^2 \, \delta\left(\langle X, t_0\rangle\right)\right] \geq \frac{\min\left(\pi, \pi^2\left(\|t_0\|_2 \sqrt{2\pi + \|t_0\|_2^2}\right)^{-1}\right)}{\sqrt{2\pi + \|t_0\|_2^2} + (\pi - 1)\|t_0\|_2} \frac{\|t\|_2^2}{8\sqrt{2\pi}}. \tag{4.54}$$

Let us now prove (4.54). We write $t = t_0^\perp + \lambda t_0$ where $t_0^\perp$ is a vector orthogonal to t_0 and $\lambda \in \mathbb{R}$. Since $\langle X, t_0^\perp \rangle$ and $\langle X, t_0 \rangle$ are independent random variables, we have

$$\begin{aligned}\mathbb{E}\left[\langle X, t\rangle^2 \, \delta\left(\langle X, t_0\rangle\right)\right] &= \mathbb{E}\left[\left\langle X, t_0^\perp\right\rangle^2\right] \mathbb{E}\left[\delta\left(\langle X, t_0\rangle\right)\right] + \lambda^2 \mathbb{E}\left[\langle X, t_0\rangle^2 \, \delta\left(\langle X, t_0\rangle\right)\right], \\ &= \left\|t_0^\perp\right\|_2^2 \mathbb{E}\delta(\|t_0\|_2 \, g) + \lambda^2 \|t_0\|_2^2 \, \mathbb{E}g^2\delta(\|t_0\|_2 \, g)\end{aligned}$$

where $g \sim \mathcal{N}(0,1)$ is standard Gaussian variable and we recall that $\delta(v) = e^v/(1+e^v)^2$ for all $v \in \mathbb{R}$. Now, it remains to lower bound $\mathbb{E}\delta(\sigma g)$ and $\mathbb{E}g^2\delta(\sigma g)$ for every $\sigma > 0$.

Since $\delta(v) \geq \exp(-|v|)/4$ for all $v \in \mathbb{R}$, one has for all $\sigma > 0$,

$$\mathbb{E}\delta(\sigma g) \geq \mathbb{E}\exp(-\sigma|g|)/4 = \exp(\sigma^2/2)\mathbb{P}[g \geq \sigma]/2$$

and

$$\mathbb{E}g^2\delta(\sigma g) \geq \mathbb{E}g^2 \exp(-\sigma|g|)/4 = (1/2)\exp(\sigma^2/2)\left[(1+\sigma^2)\mathbb{P}[g \geq \sigma] - \frac{\sigma \exp(-\sigma^2/2)}{\sqrt{2\pi}}\right].$$

Therefore, for $\sigma = \|t_0\|_2$,

$$\begin{aligned}\mathbb{E}\left[\langle X, t\rangle^2 \, \delta\left(\langle X, t_0\rangle\right)\right] \geq &\exp(\sigma^2/2)\mathbb{P}[g \geq \sigma] \left\|t_0^\perp\right\|_2^2 \\ &+ 2\lambda^2 \|t_0\|_2^2 \exp(\sigma^2/2)\left[\left(1+\sigma^2\right)\mathbb{P}[g \geq \sigma] - \frac{\sigma \exp(-\sigma^2/2)}{\sqrt{2\pi}}\right]\end{aligned}$$

and since $\|t\|_2^2 = \left\|t_0^\perp\right\|_2^2 + \lambda^2 \|t_0\|_2^2$, one has,

$$\mathbb{E}\left[\langle X,t\rangle^2 \delta\left(\langle X,t_0\rangle\right)\right] \geq \frac{\|t\|_2^2}{\sqrt{2\pi}} \min\left\{\left(\frac{1-\Phi(\sigma)}{\phi(\sigma)}\right), (1+\sigma^2)\left(\frac{1-\Phi(\sigma)}{\phi(\sigma)}\right) - \sigma\right\} \tag{4.55}$$

where ϕ and Φ denote the standard Gaussian density and distribution functions, respectively.

We lower bound the right-hand side of (4.55) using estimates on the Mills ratio $(1-\Phi)/\phi$ that follows from Equation (10) in [Dümbgen, 2010]: for every $\sigma > 0$,

$$\frac{1-\Phi(\sigma)}{\phi(\sigma)} > \frac{\pi}{\sqrt{2\pi+\sigma^2}+(\pi-1)\sigma}.$$

■

4.11.2 Proof of Section 4.6.3

Proof of Proposition 4.4: We globally follow a proof of [Elsener and van de Geer, 2016]. We have

$$P\mathcal{L}_f = \mathbb{E}[\rho_\tau(Y-f(X)) - \rho_\tau(Y-f^*(X))] = \mathbb{E}\Big\{\mathbb{E}[\rho_\tau(Y-f(X)) - \rho_\tau(Y-f^*(X))|X]\Big\}.$$

For all $x \in \mathcal{X}$, denote by F_x the c.d.f. associated with f_x. We have

$$\mathbb{E}[\rho_\tau(Y-f(X))|X=x] = (\tau-1)\int_{y<f(x)} (y-f(x))F_x(\mathrm{d}y) + \tau\int_{y\geq f(x)} (y-f(x))F_x(\mathrm{d}y)$$

$$= \int_{y\geq f(x)} (y-f(x))F_x(\mathrm{d}y) + (\tau-1)\int_{\mathbb{R}} (y-f(x))F_x(\mathrm{d}y)$$

$$= \int_{y\geq f(x)} (1-F_x(y))\mathrm{d}y + (\tau-1)\left(\int_{\mathbb{R}} yF_x(\mathrm{d}y) - f(x)\right) = g(x,f(x)) + (\tau-1)\int_{\mathbb{R}} yF_x(\mathrm{d}y)$$

where $g(x,a) = \int_{y\geq a}(1-F_x(y))\mathrm{d}y + (1-\tau)a$. Note that $\partial_2 g(x,f^*(x)) = 0$ (can be checked by calculations but also obvious from the definition). So

$$\mathbb{E}[\rho_\tau(Y-f(X)) - \rho_\tau(Y-f^*(X))|X=x] = g(x,f(x)) - g(x,f^*(x)) = \int_{f^*(x)}^{f(x)} (f(x)-u)\partial_2^2 g(x,u)\mathrm{d}u$$

$$= \int_{f^*(x)}^{f(x)} (f(x)-u)f_x(u)\mathrm{d}u \geq \frac{1}{C}\int_{f^*(x)}^{f(x)} (f(x)-u)\mathrm{d}u = \frac{(f(x)-f^*(x))^2}{2C^2}.$$

It follows that

$$\mathcal{E}_{quantile}(f) = P\mathcal{L}_f \geq \mathbb{E}\left\{\frac{(f(X)-f^*(X))^2}{2C}\right\} = \frac{1}{2C}\|f-f^*\|_{L_2}^2.$$

■

4.12 Technical lemmas

Lemma 4.4. *If $\rho \to r(2\rho)/\rho$ is non-increasing then $\rho \to \Delta(\rho)/\rho$ is non-decreasing.*

Proof. We have for all $\rho > 0$

$$\frac{\Delta(\rho)}{\rho} = \inf_{H \in S \cap (r(2\rho)/\rho) B_{L_2}} \sup_{G \in \partial \|\cdot\|(M^*)} \langle H, G \rangle .$$

The result follows since $\rho \to S \cap (r(2\rho)/\rho) B_{L_2}$ is non-increasing. ■

Lemma 4.5. *Let $\rho > 0$. The function $h : r > 0 \to w(\rho B \cap r B_{L_2})/r$ is non-increasing.*

Proof. Let $r_1 \geq r_2$. By convexity of B and B_{L_2}, we have

$$(\rho B \cap r_1 B_{L_2})/r_1 = (\rho/r_1) B \cap B_{L_2} \subset (\rho/r_2) B \cap B_{L_2} = (\rho B \cap r_2 B_{L_2})/r_2. \tag{4.56}$$

■

Chapter 5

Divide and conquer in ABC: Expectation-Progagation algorithms for likelihood-free inference

Joint work with Simon Barthelmé and Nicolas Chopin. To be published in Handbook of Approximate Bayesian Computation.

Abstract

ABC algorithms are notoriously expensive in computing time, as they require simulating many complete artificial datasets from the model. We advocate in this paper a "divide and conquer" approach to ABC, where we split the likelihood into n factors, and combine in some way n 'local' ABC approximations of each factor. This has two advantages: (a) such an approach is typically much faster than standard ABC; and (b) it makes it possible to use local summary statistics (i.e. summary statistics that depend only on the data-points that correspond to a single factor), rather than global summary statistics (that depend on the complete dataset).

This greatly alleviates the bias introduced by summary statistics, and even removes it entirely in situations where local summary statistics are simply the identity function.

We focus on EP (Expectation-Propagation), a convenient and powerful way to combine n local approximations into a global approximation. Compared to the EP-ABC approach of Barthelmé and Chopin [2014], we present two variations; one based on the parallel EP algorithm of Cseke and Heskes [2011], which has the advantage of being implementable on a parallel architecture; and one version which bridges the gap between standard EP and parallel EP. We illustrate our approach with an expensive application of ABC, namely inference on spatial extremes.

5.1 Introduction

A standard ABC algorithm samples in some way from the pseudo-posterior:

$$p_\epsilon^{\text{std}}(\boldsymbol{\theta}|\mathbf{y}^*) \propto p(\boldsymbol{\theta}) \int p(\mathbf{y}|\boldsymbol{\theta}) \mathbb{I}_{\{\|s(\mathbf{y})-s(\mathbf{y}^\star)\| \leq \epsilon\}} \, d\mathbf{y} \tag{5.1}$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ denotes the likelihood of data $\boldsymbol{y} \in \mathcal{Y}$ given parameter $\boldsymbol{\theta} \in \Theta$, $\mathbf{y}^*$ is the actual data, s is some function of the data called a 'summary statistic', and $\epsilon > 0$. As discussed elsewhere in this book, there are various ways to sample from (5.1), e.g. rejection, MCMC [Marjoram et al., 2003], SMC [Sisson et al., 2007, Beaumont et al., 2009, Del Moral et al., 2012], etc., but they all require simulating a large number of complete datasets $\boldsymbol{y}^j$ from the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, for different values of $\boldsymbol{\theta}$. This is typically the bottleneck of the computation. Another drawback of standard ABC is the dependence on s: as $\epsilon \to 0$, $p_\epsilon^{\text{std}}(\boldsymbol{\theta}|\mathbf{y}^*) \to p(\boldsymbol{\theta}|s(\boldsymbol{y}^\star)) \neq p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$, the true posterior distribution, and there is no easy way to choose s such that $p(\boldsymbol{\theta}|s(\boldsymbol{y}^\star)) \approx p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$.

In this paper, we assume that the data may be decomposed into n "chunks", $\boldsymbol{y} = (y_1, \ldots, y_n)$, and that the likelihood may be factorised accordingly:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})$$

in such a way that it is possible to sample pseudo-data y_i from each factor

$f_i(y_i|\boldsymbol{\theta})$. The objective is to approximate the pseudo-posterior:

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int f_i(y_i|\boldsymbol{\theta}) \mathbb{I}_{\left\{ \|s_i(y_i) - s_i(y_i^\star)\| \leq \epsilon \right\}} \, dy_i \right\}$$

where s_i is a "local" summary statistic, which depends only on y_i. We expect the bias introduced by the n local summary statistics s_i to be much smaller than the bias introduced by the global summary statistic s. In fact, there are practical cases where we may take $s_i(y_i) = y_i$, removing this bias entirely.

Note that we do not restrict to models such that the chunks y_i are independent. In other words, we allow each factor f_i to implicitly depends on other data-points. For instance, we could have a Markov model, with $f_i(y_i|\boldsymbol{\theta}) = p(y_i|y_{i-1}, \boldsymbol{\theta})$, or even a model with a more complicated dependence structure, say $f_i(y_i|\boldsymbol{\theta}) = p(y_i|y_{1:i-1}, \boldsymbol{\theta})$. The main requirement, however, is that we are able to sample from each factor $f_i(y_i|\boldsymbol{\theta})$. For instance, in the Markov case, this means we are able to sample from the model realisations of variable y_i, conditional on $y_{i-1} = y_{i-1}^\star$ and $\boldsymbol{\theta}$.

Alternatively, in cases where the likelihood does not admit a simple factorisation, one may replace it by some factorisable pseudo-likelihood; e.g. a marginal composite likelihood:

$$p^{\mathrm{MCL}}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

where $p(y_i|\boldsymbol{\theta})$ is the marginal density of variable y_i. Then one would take $f_i(y_i|\boldsymbol{\theta}) = p(y_i|\boldsymbol{\theta})$ (assuming we are able to simulate from the marginal distribution of y_i). Conditional distributions may be used as well; see Varin et al. [2011] for a review of composite likelihoods. Of course, replacing the likelihood by some factorisable pseudo-likelihood adds an extra level of approximation, and one must determine in practice whether the computational benefits are worth the extra cost. Estimation based on composite likelihoods is generally consistent, but their use in a Bayesian setting results in posterior distributions that are overconfident (the variance is too small, as dependent data are effectively treated as independent observations).

Many authors have taken advantage of factorisations to speed up ABC. ABC strategies for hidden Markov models are discussed in Dean et al. [2014] and Yıldırım et al. [2014]; see the review of [Jasra, 2015]. White et al. [2015] describe a method based on averages of pseudo-posteriors, which in the Gaussian case reduces to just doing one pass

of parallel EP. Ruli et al. [2015] use composite likelihoods to define low-dimensional summary statistics.

We focus on Expectation Propagation (EP, Minka, 2001), a widely successful algorithm for variational inference. In Barthelmé and Chopin [2014], we showed how to adapt EP to a likelihood-free setting. Here we extend this work with a focus on a parallel variant of EP [Cseke and Heskes, 2011] that enables massive parallelisation of ABC inference. For textbook descriptions of EP, see e.g. Section 10.7 of [Bishop, 2006] or Section 13.8 of [Gelman et al., 2014].

The chapter is organised as follows. Section 5.2 gives a general presentation of both sequential and parallel EP algorithms. Section 5.3 explains how to adapt these EP algorithms to ABC contexts. It discusses in particular some ways to speed up EP-ABC. Section 5.4 discusses how to apply EP-ABC to spatial extreme models. Section 5.5 concludes.

We use the following notations throughout: bold symbols refer to vectors or matrices, e.g. $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$. For data-points, we use (bold) $\boldsymbol{y}$ to denote complete datasets, and y_i to denote data "chunks", although we do not necessarily assume the y_i's to be scalars. The letter p typically refers to probability densities relative to the model: $p(\boldsymbol{\theta})$ is the prior, $p(y_1|\boldsymbol{\theta})$ is the likelihood of the first data chunk, and so on. The transpose of matrix $\mathbf{A}$ is denoted $\mathbf{A}^t$.

5.2 EP algorithms

5.2.1 General presentation

Consider a posterior distribution $\pi(\boldsymbol{\theta})$ that may be decomposed into $(n+1)$ factors:

$$\pi(\boldsymbol{\theta}) \propto \prod_{i=0}^{n} l_i(\boldsymbol{\theta})$$

where, say, $l_0(\boldsymbol{\theta})$ is the prior, and $l_1, \ldots, l_n$ are n contributions to the likelihood. Expectation-Progagation [EP, Minka, 2001] approximates π by a similar decomposition

$$q(\boldsymbol{\theta}) \propto \prod_{i=0}^{n} q_i(\boldsymbol{\theta})$$

where each 'site' q_i is updated in turn, conditional on the other factors, in a spirit close to a coordinate-descent algorithm.

To simplify this rather general framework, one often assumes that the q_i belong to some exponential family of distributions $\mathcal{Q}$ [Seeger, 2005]:

$$q_i(\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\lambda}_i^t \boldsymbol{t}\left(\boldsymbol{\theta}\right) - \phi\left(\boldsymbol{\lambda}_i\right)\right\}$$

where $\boldsymbol{\lambda}_i \in \mathbb{R}^d$ is the natural parameter, $\boldsymbol{t}\left(\boldsymbol{\theta}\right)$ is some function $\Theta \to \mathbb{R}^d$, and ϕ is known variously as the *log-partition function* or the *cumulant function*: $\phi(\boldsymbol{\lambda}) = \log\left[\int \exp\left\{\boldsymbol{\lambda}^t \boldsymbol{t}\left(\boldsymbol{\theta}\right)\right\} d\boldsymbol{\theta}\right]$. Working with exponential families is convenient for a number of reasons. In particular, the global approximation q is automatically in the same family, and with parameter $\boldsymbol{\lambda} = \sum_{i=0}^{n} \boldsymbol{\lambda}_i$:

$$q(\boldsymbol{\theta}) \propto \exp\left\{\left(\sum_{i=0}^{n} \boldsymbol{\lambda}_i\right)^t \boldsymbol{t}(\boldsymbol{\theta})\right\}.$$

The next section gives additional properties of exponential families upon which EP relies. Then Section 5.2.3 explains how to perform a site update, that is, how to update $\boldsymbol{\lambda}_i$, conditional on the $\boldsymbol{\lambda}_j$, $j \neq i$, so as, informally, to make q progressively closer and closer to π.

5.2.2 Properties of exponential families

Let $\mathrm{KL}(\pi||q)$ be the Kullback-Leibler divergence of q from π:

$$\mathrm{KL}(\pi||q) = \int \pi(\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\, d\boldsymbol{\theta}.$$

For a generic member $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\lambda}^t \boldsymbol{t}(\boldsymbol{\theta}) - \phi(\boldsymbol{\lambda})\right\}$ of our exponential family $\mathcal{Q}$, we have:

$$\frac{d}{d\boldsymbol{\lambda}} \mathrm{KL}(\pi||q_\lambda) = \frac{d}{d\boldsymbol{\lambda}} \phi\left(\boldsymbol{\lambda}\right) - \int \pi(\boldsymbol{\theta}) \boldsymbol{t}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{5.2}$$

where the derivative of the partition function may be obtained as:

$$\frac{d}{d\boldsymbol{\lambda}} \phi\left(\boldsymbol{\lambda}\right) = \int \boldsymbol{t}(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\lambda}^t \boldsymbol{t}\left(\boldsymbol{\theta}\right) - \phi(\boldsymbol{\lambda})\right\} d\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\lambda}}\left\{\boldsymbol{t}(\boldsymbol{\theta})\right\}. \tag{5.3}$$

Let $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\lambda}}\left\{\boldsymbol{t}(\boldsymbol{\theta})\right\}$; $\boldsymbol{\eta}$ is called the moment parameter, and there is a one-to-one correspondence between $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$; abusing notations, if $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\lambda})$ then $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\eta})$. One may interpret (5.2) as follows: finding the $q_{\boldsymbol{\lambda}}$ closest to π (in the Kullback-Leibler sense) amounts to perform

moment matching, that is, to set $\boldsymbol{\lambda}$ such that the expectation of $\boldsymbol{t}(\boldsymbol{\theta})$ under π and under $q_{\boldsymbol{\lambda}}$ match.

To make this discussion more concrete, consider the Gaussian case:

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{Q}\boldsymbol{\theta} + \boldsymbol{r}^t\boldsymbol{\theta}\right\}, \quad \boldsymbol{\lambda} = \left(\boldsymbol{r}, -\frac{1}{2}\boldsymbol{Q}\right), \quad t(\boldsymbol{\theta}) = \left(\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^t\right)$$

and the moment parameter is $\boldsymbol{\eta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^t)$, with $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$, $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{r}$. (More precisely, $\boldsymbol{\theta}^t\boldsymbol{Q}\boldsymbol{\theta} = \text{trace}(\boldsymbol{Q}\boldsymbol{\theta}\boldsymbol{\theta}^t) = \text{vect}(\boldsymbol{Q})^t\text{vect}(\boldsymbol{\theta}\boldsymbol{\theta}^t)$, so the second component of $\boldsymbol{\lambda}$ (respectively $\boldsymbol{t}(\boldsymbol{\theta})$) should be $-(1/2)\text{vect}(\boldsymbol{Q})$ (resp. $\text{vect}(\boldsymbol{\theta}\boldsymbol{\theta}')$). But, for notational convenience, our derivations will be in terms of matrices $\boldsymbol{Q}$ and $\boldsymbol{\theta}\boldsymbol{\theta}'$, rather than their vectorised versions.)

In the Gaussian case, minimising $\text{KL}(\pi||q_{\boldsymbol{\lambda}})$ amounts to take $\boldsymbol{\lambda}$ such that the corresponding moment parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^t)$ is such that $\boldsymbol{\mu} = \mathbb{E}_\pi[\boldsymbol{\theta}]$, $\boldsymbol{\Sigma} = \text{Var}_\pi[\boldsymbol{\theta}]$. We will focus on the Gaussian case in this paper (i.e. EP computes iteratively a Gaussian approximation of π), but we go on with the more general description of EP in terms of exponential families, as this allows for more compact notations, and also because we believe that other approximations could be useful in the ABC context.

5.2.3 Site update

We now explain how to perform a site update for site i, that is, how to update given $\boldsymbol{\lambda}_i$, assuming $(\boldsymbol{\lambda}_j)_{j\neq i}$ is fixed. Consider the 'hybrid' distribution:

$$\begin{aligned} h(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta})\frac{l_i(\boldsymbol{\theta})}{q_i(\boldsymbol{\theta})} &= l_i(\boldsymbol{\theta})\prod_{j\neq i} q_j(\boldsymbol{\theta}) \\ &= l_i(\boldsymbol{\theta})\exp\left\{\left(\sum_{j\neq i}\boldsymbol{\lambda}_j\right)^t \boldsymbol{t}(\boldsymbol{\theta})\right\}; \end{aligned}$$

that is, h is obtained by replacing site q_i by the true factor l_i in the global approximation q. The hybrid can be viewed as a "pseudo-posterior" distribution, formed of the product of a "pseudo-prior" q_i and a single likelihood site l_i. The update of site i is performed by minimising $\text{KL}(h||q)$ with respect to $\boldsymbol{\lambda}_i$ (again, assuming the other $\boldsymbol{\lambda}_j$, $j \neq i$, are fixed). Informally, this may be interpreted as a local projection (in the Kullback-Leibler sense) of π to $\mathcal{Q}$.

Given the properties of exponential families laid out in the previous section, one sees that this site update amounts to setting $\boldsymbol{\lambda}_i$ so that

$\boldsymbol{\lambda} = \sum_j \boldsymbol{\lambda}_j$ matches $\mathbb{E}_h[\boldsymbol{t}(\boldsymbol{\theta})]$, the expectation of $t(\boldsymbol{\theta})$ with respect to the hybrid distribution. In addition, one may express the update of $\boldsymbol{\lambda}_i$ as a function of the current values of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}$, using the fact that $\sum_{j \neq i} \boldsymbol{\lambda}_j = \boldsymbol{\lambda} - \boldsymbol{\lambda}_i$, as done below in Algorithm 5.1.

Algorithm 5.1 Generic site update in EP

Function SiteUpdate$(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$:
1. Compute

$$\boldsymbol{\lambda}^{\text{new}} := \boldsymbol{\lambda}\left(\mathbb{E}_h[\boldsymbol{t}(\boldsymbol{\theta})]\right), \quad \boldsymbol{\lambda}_i^{\text{new}} := \boldsymbol{\lambda}^{\text{new}} - \boldsymbol{\lambda} + \boldsymbol{\lambda}_i$$

where $\boldsymbol{\eta} \to \boldsymbol{\lambda}(\boldsymbol{\eta})$ is the function that maps the moment parameters to the natural parameters (for the considered exponential family, see previous section) and

$$\mathbb{E}_h\left[\boldsymbol{t}(\boldsymbol{\theta})\right] = \frac{\int \boldsymbol{t}(\boldsymbol{\theta}) l_i(\boldsymbol{\theta}) \exp\left\{(\boldsymbol{\lambda} - \boldsymbol{\lambda}_i)^t \boldsymbol{t}(\boldsymbol{\theta})\right\} d\boldsymbol{\theta}}{\int l_i(\boldsymbol{\theta}) \exp\left\{(\boldsymbol{\lambda} - \boldsymbol{\lambda}_i)^t \boldsymbol{t}(\boldsymbol{\theta})\right\} d\boldsymbol{\theta}}. \tag{5.4}$$

2. Return $\boldsymbol{\lambda}_i^{\text{new}}$, and optionally $\boldsymbol{\lambda}^{\text{new}}$ (as determined by syntax, i.e. either $\boldsymbol{\lambda}_i^{\text{new}} \leftarrow$ SiteUpdate$(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$, or $(\boldsymbol{\lambda}_i^{\text{new}}, \boldsymbol{\lambda}^{\text{new}}) \leftarrow$ SiteUpdate$(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$).

In practice, the feasibility of EP for a given posterior is essentially determined by the difficulty to evaluate, or approximate, the integral (5.4). Note the simple interpretation of this quantity: this is the posterior expectation of $\boldsymbol{t}(\boldsymbol{\theta})$, for pseudo-prior q_{-i}, and pseudo-likelihood the likelihood factor $l_i(\boldsymbol{\theta})$. (In the EP literature, the pseudo-prior q_{-i} is often called the cavity distribution, and the pseudo-posterior $\propto q_{-i}(\boldsymbol{\theta}) l_i(\boldsymbol{\theta})$ the tilted or hybrid distribution.)

5.2.4 Gaussian sites

In this paper, we will focus on Gaussian approximations; that is $\mathcal{Q}$ is the set of Gaussian densities

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q} \boldsymbol{\theta} + \boldsymbol{r}^t \boldsymbol{\theta}\right\}, \quad \boldsymbol{\lambda} = \left(\boldsymbol{r}, -\frac{1}{2}\boldsymbol{Q}\right)$$

and EP computes iteratively a Gaussian approximation of π, obtained as a product of Gaussian factors. For this particular family, simple calculations show that the site updates take the form given by Algorithm 5.2.

In words, one must compute the expectation and variance of the pseudo-posterior obtained by multiplying the Gaussian pseudo-prior q_{-i}, and likelihood l_i.

Algorithm 5.2 EP Site update (Gaussian case)

Function SiteUpdate$(i, l_i, (\boldsymbol{r}_i, \boldsymbol{Q}_i), (\boldsymbol{r}, \boldsymbol{Q}))$:

1. Compute

$$
\begin{aligned}
Z_h &= \int q_{-i}(\boldsymbol{\theta}) l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
\boldsymbol{\mu}_h &= \frac{1}{Z_h} \int \boldsymbol{\theta} q_{-i}(\boldsymbol{\theta}) l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
\boldsymbol{\Sigma}_h &= \frac{1}{Z_h} \int \boldsymbol{\theta}\boldsymbol{\theta}^t q_{-i}(\boldsymbol{\theta}) l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} - \boldsymbol{\mu}_h \boldsymbol{\mu}_h^t
\end{aligned}
$$

   where $q_{-i}(\boldsymbol{\theta})$ is the Gaussian density

$$
q_{-i}(\boldsymbol{\theta}) \propto \exp\left\{ -\frac{1}{2} \boldsymbol{\theta}^t \left(\boldsymbol{Q} - \boldsymbol{Q}_i\right) \boldsymbol{\theta} + \left(\boldsymbol{r} - \boldsymbol{r}_i\right)^t \boldsymbol{\theta} \right\}.
$$

2. Return $(\boldsymbol{r}_i^{\text{new}}, \boldsymbol{Q}_i^{\text{new}})$, and optionally $(\boldsymbol{r}^{\text{new}}, \boldsymbol{Q}^{\text{new}})$ (according to syntax as in Algorithm 5.1), where

$$
\begin{aligned}
(\boldsymbol{Q}^{\text{new}}, \boldsymbol{r}^{\text{new}}) &= \left(\boldsymbol{\Sigma}_h^{-1}, \boldsymbol{\Sigma}_h^{-1} \boldsymbol{\mu}_h\right), \\
(\boldsymbol{Q}_i^{\text{new}}, \boldsymbol{r}_i^{\text{new}}) &= \left(\boldsymbol{Q}_i + \boldsymbol{Q}^{\text{new}} - \boldsymbol{Q}, \boldsymbol{r}_i + \boldsymbol{r}^{\text{new}} - \boldsymbol{r}\right).
\end{aligned}
$$

5.2.5 Order of site updates: sequential EP, parallel EP, and block-parallel EP

We now discuss in which *order* the site updates may be performed; i.e. should site updates be performed sequentially, or in parallel, or something in between.

The initial version of EP, as described in Minka [2001], was purely sequential (and will therefore be referred to as "sequential EP" from now on): one updates $\boldsymbol{\lambda}_0$ given the current values of $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n$, then one updates $\boldsymbol{\lambda}_1$ given $\boldsymbol{\lambda}_0$ (as modified in the previous update) and $\boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_n$, and so on; see Algorithm 5.3. Since the function SiteUpdate $(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$ computes the updated version of both $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda} = \sum_{j=0}^n \boldsymbol{\lambda}_j$, $\boldsymbol{\lambda}$ changes at each call of SiteUpdate.

Algorithm 5.3 is typically run until $\boldsymbol{\lambda} = \sum_{i=0}^n \boldsymbol{\lambda}_i$ stabilises in some sense.

The main drawback of sequential EP is that, given its sequential na-

Algorithm 5.3 Sequential EP

Require: initial values for $\boldsymbol{\lambda}_0, \ldots, \boldsymbol{\lambda}_n$
  $\boldsymbol{\lambda} \leftarrow \sum_{i=0}^{n} \boldsymbol{\lambda}_i$
  repeat
    for $i = 0$ to n **do**
      $(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}) \leftarrow \text{SiteUpdate}\,(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$
    end for
  until convergence
  return $\boldsymbol{\lambda}$

ture, it is not easily amenable to parallel computation. Cseke and Heskes [2011] proposed a parallel EP algorithm, where all sites are updated in parallel, independently of each other. This is equivalent to update the sum $\boldsymbol{\lambda} = \sum_{i=0}^{n} \boldsymbol{\lambda}_i$ only after all the sites have been updated; see Algorithm 5.4.

Algorithm 5.4 Parallel EP

Require: initial values for $\boldsymbol{\lambda}_0, \ldots, \boldsymbol{\lambda}_n$
  $\boldsymbol{\lambda} \leftarrow \sum_{i=0}^{n} \boldsymbol{\lambda}_i$
  repeat
    for $i = 0$ to n **do** (parallel)
      $\boldsymbol{\lambda}_i \leftarrow \text{SiteUpdate}\,(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$
    end for
    $\boldsymbol{\lambda} \leftarrow \sum_{i=0}^{n} \boldsymbol{\lambda}_i$
  until convergence
  return $\boldsymbol{\lambda}$

Parallel EP is "embarrassingly parallel", since its inner loop performs $(n + 1)$ independent operations. A drawback of parallel EP is that its convergence is typically slower (i.e. requires more complete passes over all the sites) than sequential EP. Indeed, during the first pass, all the sites are provided with the same initial global approximation $\boldsymbol{\lambda}$, whereas in sequential EP, the first site updates allow to refine progressively $\boldsymbol{\lambda}$, which makes the following updates easier.

We now propose a simple hybrid of these two EP algorithms, which we call block-parallel EP. We assume we have n_{core} cores (single processing units) at our disposal. For each block of n_{core} successive sites, we update these n_{core} sites in parallel, and then update the global approximation $\boldsymbol{\lambda}$ after these n_{core} updates; see Algorithm 5.5.

Algorithm 5.5 Block-parallel EP

Require: initial values for $\boldsymbol{\lambda}_0, \ldots, \boldsymbol{\lambda}_n$
  $\boldsymbol{\lambda} \leftarrow \sum_{i=0}^{n} \boldsymbol{\lambda}_i$
  repeat
    for $k = 1$ to $\lceil (n+1)/n_{\text{core}} \rceil$ **do**
      for $i = (k-1)n_{\text{core}}$ to $(kn_{\text{core}} - 1) \wedge n$ **do** (parallel)
        $\boldsymbol{\lambda}_i \leftarrow \text{SiteUpdate}\,(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$
      end for
      $\boldsymbol{\lambda} \leftarrow \sum_{i=0}^{n} \boldsymbol{\lambda}_i$
    end for
  until convergence
  return $\boldsymbol{\lambda}$

Quite clearly, block-parallel EP generalises both sequential EP (take $n_{\text{core}} = 1$) and parallel EP (take $n_{\text{core}} = n + 1$). This generalisation is useful in any situation where the actual number of cores n_{core} available in a given architecture is such that $n_{\text{core}} \ll (n+1)$. In this way, we achieve essentially the same speed-up as Parallel EP in terms of parallelisation (since only n_{core} cores are available anyway), but we also progress faster thanks to the sequential nature of the successive block updates. We shall discuss more specifically in the next section the advantage of block-parallel EP over standard parallel EP in an ABC context.

5.2.6 Other practical considerations

Often, the prior, which was identified with l_0 in our factorisation, already belongs to the approximating parametric family: $p(\boldsymbol{\theta}) = q_{\boldsymbol{\lambda}_0}(\boldsymbol{\theta})$. In that case, one may fix beforehand $q_0(\boldsymbol{\theta}) = l_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, and update only $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n$ in the course of the algorithm, while keeping $\boldsymbol{\lambda}_0$ fixed to the value given by the prior.

EP also provides at no extra cost an approximation of the normalising constant of π: $Z = \int_{\boldsymbol{\theta}} \prod_{i=0}^{n} l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. When π is a posterior, this can be used to approximate the marginal likelihood (evidence) of the model. See e.g. Barthelmé and Chopin [2014] for more details.

In certain cases, EP updates are "too fast", in the sense that the update of difficult sites may lead to e.g. degenerate precision matrices (in the Gaussian case). One well known method to slow down EP is to perform fractional updates [Minka, 2004]; that is, informally, update only a fraction $\alpha \in (0, 1]$ of the site parameters; see Algorithm 5.6.

Algorithm 5.6 Generic site update in EP (fractional version, requires $\alpha \in (0, 1]$)

Function SiteUpdate$(i, l_i, \boldsymbol{\lambda}_i, \boldsymbol{\lambda})$:
1. Compute

$$\boldsymbol{\lambda}^{\text{new}} := \alpha\boldsymbol{\lambda}\left(\mathbb{E}_h[\boldsymbol{t}(\boldsymbol{\theta})]\right) + (1-\alpha)\boldsymbol{\lambda}, \quad \boldsymbol{\lambda}_i^{\text{new}} := \boldsymbol{\lambda}_i + \alpha\left\{\boldsymbol{\lambda}\left(\mathbb{E}_h[\boldsymbol{t}(\boldsymbol{\theta})]\right) - \boldsymbol{\lambda}\right\}$$

with $\mathbb{E}_h[\boldsymbol{t}(\boldsymbol{\theta})]$ defined in (5.3), see Step 1 of 5.1.
2. As Step 2 of Algorithm 5.1.

In practice, reducing α is often the first thing to try when EP either diverges or fails because of non-invertible matrices (in the Gaussian case). Of course, the price to pay is that with a lower α, EP may require more iterations to converge.

5.2.7 Theoretical properties of EP

EP is known to work well in practice, sometimes surprisingly so, but it has proved quite resilient to theoretical study. In Barthelmé and Chopin [2014] we could give no guarantees whatsoever, but since then the situation has improved. The most important question concerns the quality of the approximations produced by EP. Under relatively strong conditions Dehaene and Barthelmé [2015a] were able to show that Gaussian EP is asymptotically exact in the large-data limit. This means that if the posterior tends to a Gaussian (which usually happens in identifiable models), then EP will recover the exact posterior. Dehaene and Barthelmé [2015b] show further that EP recovers the mean of the posterior with an error that vanishes in $\mathcal{O}(n^{-2})$, where n is the number of data-points. The error is up to an order of magnitude lower than what one can expect from the canonical Gaussian approximation, which uses the mode of the posterior as an approximation to the mean.

However, in order to have an EP approximation, one needs to find one in the first place. The various flavours of EP (including the ones described here) are all relatively complex fixed-point iterations and their convergence is hard to study. Dehaene and Barthelmé [2015a] show that parallel EP converges in the large-data limit to a Newton iteration, and inherits the potential instabilities in Newton's method. Just like Newton's method, non-convergence in EP can be fixed by slowing down the iterations, as described above.

The general picture is that EP should work very well if the hybrids are well-behaved (log-concave, roughly). Like any Gaussian approximation it can be arbitrarily poor when used on multi-modal posterior distributions, unless the modes are all equivalent.

Note finally that the results above apply to variants of EP where hybrid distributions are tractable (meaning their moments can be computed exactly). In ABC applications that is not the case, and we will incur additional Monte Carlo error. As we will explain, part of the trick in using EP in ABC settings is finding ways of minimising that additional source of errors.

5.3 Applying EP in ABC

5.3.1 Principle

Recall that our objective is to approximate the ABC posterior

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int f_i(y_i|\boldsymbol{\theta}) \mathbb{I}_{\left\{ \left\| s_i(y_i) - s_i(y_i^\star) \right\| \leq \epsilon \right\}} \, dy_i \right\}$$

for a certain factorisation of the likelihood, and for a certain collection of local summary statistics s_i. This immediately suggests using EP on the following collection of sites

$$l_i(\boldsymbol{\theta}) = \int f_i(y_i|\boldsymbol{\theta}) \mathbb{I}_{\left\{ \left\| s_i(y_i) - s_i(y_i^\star) \right\| \leq \epsilon \right\}} \, dy_i$$

for $i = 1, \ldots, n$. For convenience, we focus on the Gaussian case (i.e. the l_i's will be approximated by Gaussian factors q_i), and assume that the prior $p(\boldsymbol{\theta})$ itself is already Gaussian, and does not need to be approximated.

From Algorithm 5.2, we see that, in this Gaussian case, it is possible to perform a site update provided that we are able to compute the mean and variance of a pseudo-posterior, corresponding to a Gaussian prior q_{-i}, and likelihood l_i.

Algorithm 5.7 describes a simple rejection algorithm that may be used to perform the site update. Using this particular algorithm inside sequential EP leads to the EP-ABC algorithm derived in Barthelmé and Chopin [2014]. We stress however that one may generally use any ABC approach to perform such a site update. The main point is that this local ABC

problem is much simpler than ABC for the complete likelihood for two reasons. First, the pseudo-prior q_{-i} is typically much more informative than the true prior $p(\boldsymbol{\theta})$, because q_{-i} approximates the posterior of all the data minus y_i. Thus, we are much less likely to sample values of $\boldsymbol{\theta}$ with low likelihood. Second, even for a fixed $\boldsymbol{\theta}$, the probability that $\|s_i(y_i) - s_i(y_i^\star)\| \leq \epsilon$ is typically much larger than $\|s(\boldsymbol{y}) - s(\boldsymbol{y}^\star)\| \leq \epsilon$, as s_i is generally of lower dimension than s.

Algorithm 5.7 Local ABC algorithm to perform site update

Function SiteUpdate$(i, f_i, (\boldsymbol{r}_i, \boldsymbol{Q}_i), (\boldsymbol{r}, \boldsymbol{Q}))$:

1. Simulate $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \sim N(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$ where $\boldsymbol{\Sigma}_{-i}^{-1} = \boldsymbol{Q} - \boldsymbol{Q}_i$, $\boldsymbol{\mu}_{-i} = \boldsymbol{\Sigma}_{-i}(\boldsymbol{r} - \boldsymbol{r}_i)$.
2. For each $m = 1, \ldots M$, simulate $y_i^{(m)} \sim f_i(\cdot|\boldsymbol{\theta}^{(m)})$.
3. Compute

$$M_{\text{acc}} = \sum_{m=1}^{M} \mathbb{I}\left\{\left\|s_i(y_i^{(m)}) - s_i(y_i^\star)\right\| \leq \epsilon\right\}$$

$$\hat{\boldsymbol{\mu}}_h = \frac{1}{M_{\text{acc}}} \sum_{m=1}^{M} \boldsymbol{\theta}^{(m)} \mathbb{I}\left\{\left\|s_i(y_i^{(m)}) - s_i(y_i^\star)\right\| \leq \epsilon\right\}$$

$$\hat{\boldsymbol{\Sigma}}_h = \frac{1}{M_{\text{acc}}} \sum_{m=1}^{M} \boldsymbol{\theta}^{(m)} \left[\boldsymbol{\theta}^{(m)}\right]^t \mathbb{I}\left\{\left\|s_i(y_i^{(m)}) - s_i(y_i^\star)\right\| \leq \epsilon\right\} - \hat{\boldsymbol{\mu}}_h \hat{\boldsymbol{\mu}}_h^t$$

4. Return $(\boldsymbol{r}_i^{\text{new}}, \boldsymbol{Q}_i^{\text{new}})$, and optionally $(\boldsymbol{r}^{\text{new}}, \boldsymbol{Q}^{\text{new}})$ (according to syntax as in Algorithm 5.1), where

$$(\boldsymbol{Q}^{\text{new}}, \boldsymbol{r}^{\text{new}}) = \left(\hat{\boldsymbol{\Sigma}}_h^{-1}, \hat{\boldsymbol{\Sigma}}_h^{-1} \hat{\boldsymbol{\mu}}_h\right),$$
$$(\boldsymbol{Q}_i^{\text{new}}, \boldsymbol{r}_i^{\text{new}}) = (\boldsymbol{Q}_i + \boldsymbol{Q}^{\text{new}} - \boldsymbol{Q}, \boldsymbol{r}_i + \boldsymbol{r}^{\text{new}} - \boldsymbol{r}).$$

5.3.2 Practical considerations

We have observed that in many problems the acceptance rate of Algorithm 5.7 may vary significantly across sites, so, instead of fixing M, the number of simulated pairs $(\boldsymbol{\theta}^{(m)}, y_i^{(m)})$, to a given value, we recommend to sample until the number of accepted pairs (i.e. the number of $(\boldsymbol{\theta}^{(m)}, y_i^{(m)})$ such that $\left\|s_i(y_i^{(m)}) - s_i(y_i^{(m)})\right\| \leq \epsilon$) equals a certain thresh-

old M_0.

Another simple way to improve EP-ABC is to generate the $\boldsymbol{\theta}^{(m)}$ using quasi-Monte Carlo for distribution $N(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$, we take $\boldsymbol{\theta}^m = \boldsymbol{\mu}_{-i} + \boldsymbol{L}\boldsymbol{\Phi}^{-1}(\boldsymbol{u}^m)$, where $\boldsymbol{\Phi}^{-1}$ is the Rosenblatt transformation (multivariate quantile function) of the unit normal distribution of dimension $\dim(\boldsymbol{\theta})$, $\boldsymbol{L}\boldsymbol{L}^{\boldsymbol{t}} = \boldsymbol{\Sigma}_{-i}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_{-i}$, and the $\boldsymbol{u}^{(m)}$ is a low-discrepancy sequence, such as the Halton sequence; see e.g. Chap. 5 in Lemieux [2009] for more background on low-discrepancy sequences and quasi-Monte Carlo.

Regarding ϵ, our practical experience is that finding a reasonable value through trial and error is typically much easier with EP-ABC than with standard ABC. This is because the y_i's are typically of much lower dimension than the complete data-set $\boldsymbol{y}$. However, one more elaborate recipe to calibrate ϵ is to run EP-ABC with a first value of ϵ, then set ϵ to the minimal value such that the proportion of simulated y_i at each site such that $\|s_i(y_i) - s_i(y_i^\star)\| \leq \epsilon$ is above, say, 5%. Then one may start over with this new value of ϵ.

Another direction suggested by Mark Beaumont in a personal communication is to correct the estimated precision matrices for bias, using formula (4) from Paz and Sánchez [2015].

5.3.3 Speeding up parallel EP-ABC in the IID case

This section considers the IID case, i.e. the model assumes that the y_i are IID (independent and identically distributed), given $\boldsymbol{\theta}$: then

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_1(y_i|\boldsymbol{\theta})$$

where f_1 denotes the common density of the y_i. In this particular case, each of the n local ABC posteriors, as described by Algorithm 5.7, will use pseudo-data from the *same* distribution (given $\boldsymbol{\theta}$). This suggests recycling these simulations across sites.

Barthelmé and Chopin [2014] proposed a recycling strategy based on sequential importance sampling. Here, we present an even simpler scheme that may be implemented when Parallel EP is used. At the start of iteration t of Parallel EP, we sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the current *global* approximation of the posterior. For each $\boldsymbol{\theta}^m$, we sample $y^{(m)} \sim f_1(y|\boldsymbol{\theta}^m)$. Then, for each site i, we can compute the first two

moments of the hybrid distribution by simply doing an importance sampling step, from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to $N(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$, which is obtained by dividing the density of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by factor q_i. Specifically, the weight function is:

$$\frac{|\boldsymbol{Q}_{-i}| \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q}_{-i}\boldsymbol{\theta} + \boldsymbol{r}_{-i}^t \boldsymbol{\theta}\right\}}{|\boldsymbol{Q}| \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q}\boldsymbol{\theta} + \boldsymbol{r}^t \boldsymbol{\theta}\right\}} = \frac{|\boldsymbol{Q} - \boldsymbol{Q}_i|}{|\boldsymbol{Q}|} \exp\left\{\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q}_i \boldsymbol{\theta} - \boldsymbol{r}_i^r \boldsymbol{\theta}\right\}$$

since $\boldsymbol{Q} = \boldsymbol{Q}_i + \boldsymbol{Q}_{-i}$, $\boldsymbol{r} = \boldsymbol{r}_i + \boldsymbol{r}_{-i}$. Note that further savings can be obtained by retaining the samples for several iterations, regenerating only when the global approximation has changed too much relative to the values used for sampling. In our implementation we monitor the drift by computing the Effective Sample Size of importance sampling from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (the distribution of the current samples) for the new global approximation $N(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$.

We summarise the so-obtained algorithm as Algorithm 5.8. Clearly, recycling allows us for a massive speed-up when the number n of sites is large, as we re-use the same set of simulated pairs $(\boldsymbol{\theta}^{(m)}, y^{(m)})$ for all the n sites. In turns, this allows us to take a larger value for M, the number of simulations, which leads to more stable results.

We have advocated Parallel EP in Section 5.2 as a way to parallelise the computations over the n sites. Given the particular structure of Algorithm 5.8, we see that it is also easy to parallelise the simulation of the M pairs $(\boldsymbol{\theta}^{(m)}, y^{(m)})$ that is performed at the start of each EP iteration; this part is usually the bottleneck of the computation. In fact, we also observe that Algorithm 5.8 performs slightly better than the recycling version of EP-ABC (as described in Barthelmé and Chopin 2014) even on a non-parallel architecture.

5.4 Application to spatial extremes

We now turn our attention to likelihood-free inference for spatial extremes, following Erhardt and Smith [2012], see also Prangle [2014].

5.4.1 Background

The data $\boldsymbol{y}$ consist of n IID observations y_i, typically observed over time, where $y_i \in \mathbb{R}^d$ represents some maximal measure (e.g. rainfall) collected at d locations x_j (e.g. in $\mathbb{R}^2$). The standard modelling approach for extremes is to assign to y_i a max-stable distribution (i.e. a distribution stable by maximisation, in the same way that Gaussians are stable by

Algorithm 5.8 Parallel EP-ABC with recycling (IID case)

Require: M (number of samples), initial values for $(\boldsymbol{r}_i, \boldsymbol{Q}_i)_{i=0,\ldots,n}$ (note $(\boldsymbol{r}_0, \boldsymbol{Q}_0)$ stays constant during the course of the algorithm, as we have assumed a Gaussian prior with natural parameter $(\boldsymbol{r}_0, \boldsymbol{Q}_0)$)
repeat
  $\boldsymbol{Q} \leftarrow \sum_{i=0}^{n} \boldsymbol{Q}_i$, $\boldsymbol{r} \leftarrow \sum_{i=0}^{n} \boldsymbol{r}_i$, $\boldsymbol{\Sigma} \leftarrow \boldsymbol{Q}^{-1}$, $\boldsymbol{\mu} \leftarrow \boldsymbol{\Sigma}\boldsymbol{r}$
  for $m = 1, \ldots, M$ **do**
    $\boldsymbol{\theta}^{(m)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
    $y^{(m)} \sim f_1(y|\boldsymbol{\theta}^{(m)})$
  end for
  for $i = 1, \ldots, n$ **do**
    for $m = 1, \ldots, M$ **do**
      $w^{(m)} \leftarrow \frac{|\boldsymbol{Q}-\boldsymbol{Q}_i|}{|\boldsymbol{Q}|} \exp\left\{\frac{1}{2}(\boldsymbol{\theta}^{(m)})^t \boldsymbol{Q}_i \boldsymbol{\theta}^{(m)} - \boldsymbol{r}_i^t \boldsymbol{\theta}^{(m)}\right\} \mathbb{I}\left\{\left\| s_i(y_i^{(m)}) - s_i(y_i^\star) \right\| \leq \epsilon\right\}$
    end for
    $\hat{Z} \leftarrow M^{-1} \sum_{m=1}^{M} w^{(m)}$
    $\hat{\boldsymbol{\mu}} \leftarrow (M\hat{Z})^{-1} \times \sum_{m=1}^{M} w^{(m)} \boldsymbol{\theta}^{(m)}$
    $\hat{\boldsymbol{\Sigma}} \leftarrow (M\hat{Z})^{-1} \times \sum_{m=1}^{M} w^{(m)} \boldsymbol{\theta}^{(m)} \left[\boldsymbol{\theta}^{(m)}\right]^t - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^t$
    $\boldsymbol{r}_i \leftarrow \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} - \boldsymbol{r}_{-i}$
    $\boldsymbol{Q}_i \leftarrow \hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{Q}_{-i}$
  end for
until Stopping rule (e.g. changes in $(\boldsymbol{r}, \boldsymbol{Q})$ have become small)

addition). In the spatial case, the vector y_i is composed of d observations of a max-stable process $x \to Y(x)$ at the d locations x_j. A general approach to defining max-stable processes is [Schlather, 2002]:

$$Y(x) = \max_k \{s_k \max(0, Z_k(x))\} \tag{5.5}$$

where $(s_k)_{k=1}^{\infty}$ is the realisation of a Poisson process over $\mathbb{R}^+$ with intensity $\Lambda(\mathrm{d}s) = \mu^{-1}s^{-2}\mathrm{d}s$ (if we view the Poisson process as producing a random set of "spikes" on the positive real line, then s_1 is the location of the first spike, s_2 the second, etc.), $(Z_k)_{k=1}^{\infty}$ is a countable collection of IID realisations of a zero-mean, unit-variance stationary Gaussian process, with correlation function $\rho(h) = \mathrm{Corr}(Z_k(x), Z_k(x'))$ for x, x' such that $\|x - x'\| = h$, and $\mu = \mathbb{E}[\max(0, Z_k(x))]$. Note that $Y(x)$ is marginally distributed according to a unit Fréchet distribution, with CDF $F(y) = \exp(-1/y)$.

As in Erhardt and Smith [2012], we will consider the following parametric Whittle-Matérn correlation function

$$\rho_{\boldsymbol{\theta}}(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{h}{c}\right)^{\nu} K_\nu\left(\frac{h}{c}\right), \quad c, \nu > 0$$

where K_ν is the modified Bessel function of the third kind. We take $\boldsymbol{\theta} = (\log \nu, \log c)$ so that $\boldsymbol{\Theta} = \mathbb{R}^2$. (We will return to this logarithmic parametrisation later.)

The main issue with spatial extremes is that, unless $d \leq 2$, the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ is intractable. One approach to estimate $\boldsymbol{\theta}$ is pairwise marginal composite likelihood [Padoan et al., 2010]. Alternatively, (5.5) suggests a simple way to simulate from $p(\boldsymbol{y}|\boldsymbol{\theta})$, at least approximately (e.g. by truncating the domain of the Poisson process to $[0, S_{\max}]$). This motivates likelihood-free inference [Erhardt and Smith, 2012].

5.4.2 Summary statistics

One issue however with likelihood-free inference for this class of models is the choice of summary statistics: Erhardt and Smith [2012] compare several choices, and find that the one that performs best is some summary of the clustering of the $d(d-1)(d-2)/6$ triplet-wise coefficients

$$\frac{n}{\sum_{i=1}^{n} \{\max(y_i(x_j), y_i(x_k), y_i(x_l))\}^{-1}}, \quad 1 \leq j < k < l \leq d.$$

But computing these coefficients require $\mathcal{O}(d^3)$ operations, and may actually be more expensive than simulating the data itself: Prangle [2014] observes in a particular experiment than the cost of computing these coefficients is already more that twice the cost of simulating data for $d = 20$. As a result, the overall approach of Erhardt and Smith [2012] may take several days to run on a single-core computer.

In contrast, EP-ABC allows us to define local summary statistics, $s_i(y_i)$, that depend only on one data-point y_i. We simply take $s_i(y_i)$ to be the (2-dimensional) OLS (ordinary least squares) estimate of regression

$$\log |F(y_i(x_j)) - F(y_i(x_k))| = a + b \log \|x_j - x_k\| + \epsilon_{jk}, \quad 1 \leq j < k \leq d$$

where F is the unit Fréchet CDF. The madogram function

$$h \to \mathbb{E}\left[|Y(x) - Y(x')|\right], \text{ for } \|x - x'\| = h,$$

or its empirical version, is a common summary of spatial dependencies (for extremes). Here, we take the $F-$madogram, i.e. $Y(x)$ is replaced by $F(Y(x)) \sim U[0,1]$, because $Y(x)$ is Fréchet and thus $\mathbb{E}\left[|Y(x)|\right] = +\infty$.

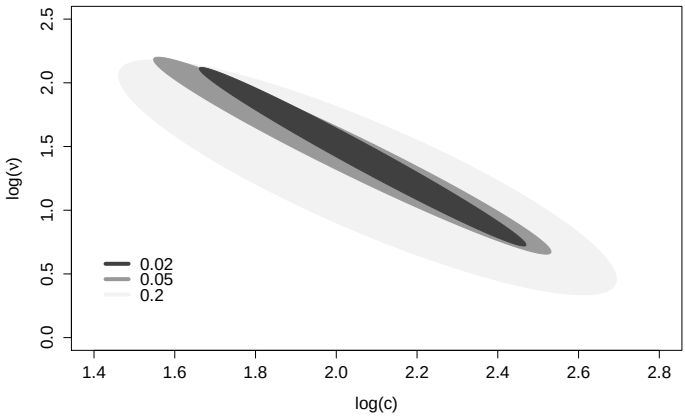


Figure 5.1 – 50% credible ellipses of the Gaussian approximation of the posterior computed by EP-ABC, for different values of ϵ, and rainfall dataset.

5.4.3 Numerical results on real data

We now apply EP-ABC to the rainfall dataset of the SpatialExtremes R package (available at `http://spatialextremes.r-forge.r-project.org/`), which records maximum daily rainfall amounts over the years 1962–2008 occurring during June–August at 79 sites in Switzerland. We ran sequential EP with recycling and quasi-Monte Carlo (see discussion in Section 5.3.2). Figure 5.1 plots the EP-ABC posterior for $\epsilon = 0.2$, 0.05 and 0.02. A $N(0, 1)$ prior was used for both components of $\boldsymbol{\theta} = (\log \nu, \log c)$.

Each run took about 3 hours on our desktop computer, and generated about 10^5 data-points (i.e. realisations $y_i \in \mathbb{R}^d$, where d is the number of stations). As a point of comparison, we ran Erhardt and Smith [2012]'s R package for a week on the same computer, which led to the generation of 5×10^4 complete datasets (i.e. $\approx 4 \times 10^6$ data-points). However, the ABC posterior approximation obtained from the 100 generated datasets that were closest to the data, relative to their summary statistics, was not significantly different from the prior.

Finally, we discuss the strong posterior correlations between the two parameters that are apparent in Figure 5.1. Figure 5.2 plots a heat map of functions $(\nu, c) \to \int |\rho_{\nu,c} - \rho_{\nu_0,c_0}|$ and $(\log \nu, \log c) \to \int |\rho_{\nu,c} - \rho_{\nu_0,c_0}|$,

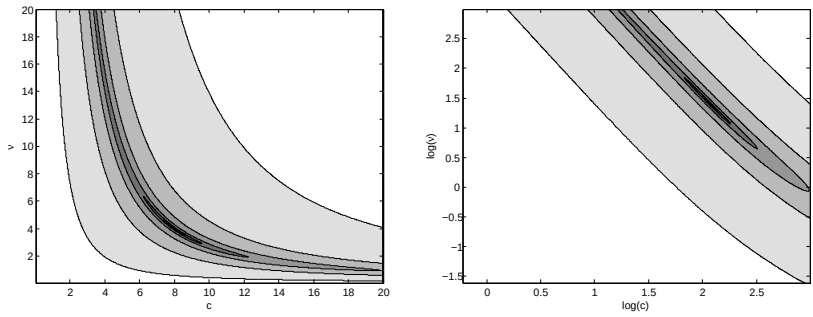


Figure 5.2 – Heat map of functions $(\nu, c) \to \int |\rho_{\nu,c} - \rho_{\nu_0,c_0}|$ and $(\log \nu, \log c) \to \int |\rho_{\nu,c} - \rho_{\nu_0,c_0}|$, for $(\nu_0, c_0) = (8, 4)$.

for $(\nu_0, c_0) = (8, 4)$. The model appears to be nearly non-identifiable, as values of (ν, c) that are far away may produce correlation functions that are nearly indistinguishable. In addition, the parametrisation $\boldsymbol{\theta} = (\log \nu, \log c)$ has the advantage of giving an approximately Gaussian shape to contours, which is clearly helpful in our case given that EP-ABC generates a Gaussian approximation. Still, it is interesting to note that EP-ABC performs well on such a nearly non-identifiable problem.

5.4.4 EP Convergence

Finally, we compare the convergence (relative to the number of iterations) of the standard version, and the block-parallel version (described in Section 5.2.5) of EP-ABC, on the rainfall dataset discussed above. Figure 5.3 plots the evolution of the posterior mean of both parameters ν (left panel) and c (right panel), relative to the number of site updates, for 3 runs of both versions, and for $\epsilon = 0.05$.

We took $n_{\text{core}} = 10$ (i.e. blocks of 10 sites are updated in parallel), although both algorithms were run on a single core. We see that both algorithms essentially converge at the same rate. Thus, if implemented on a 10-core machine, the block-parallel version should offer essentially a $\times 10$ speed-up.

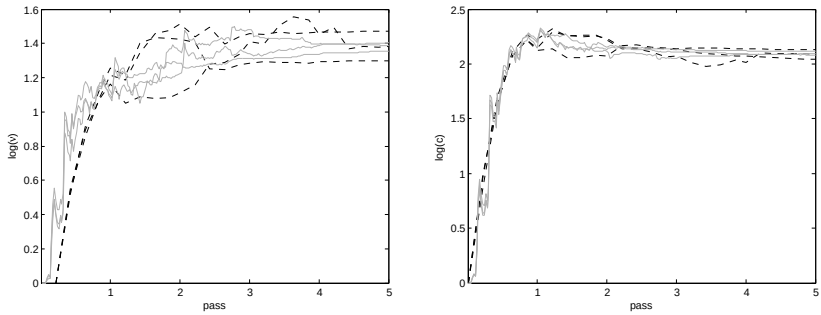


Figure 5.3 – Posterior mean of $\log \nu$ (left panel) and $\log c$ (right panel) as a function of the number of passes (one pass equals $n = 47$ site updates), for 3 runs of the sequential version (solid grey line), and block-parallel version ($n_{\text{core}} = 10$, dashed black line) of EP-ABC, applied to rainfall dataset ($\epsilon = 0.05$).

5.5 Conclusion

Compared to standard ABC, the main drawback of EP-ABC is that it introduces an extra level of approximation, because of its EP component. On the other hand, EP-ABC strongly reduces, or sometimes removes entirely, the bias introduced by summary statistics, as it makes possible to use n local summaries, instead of just one for the complete dataset. In our experience (see e.g. the examples in Barthelmé and Chopin [2014]), this bias reduction more than compensates the bias introduced by EP. But the main advantage of EP-ABC is that it is much faster than standard ABC. Speed-ups of more than 100 are common, as evidenced by our spatial extremes example.

We have developed a Matlab package, available at `https://sites.google.com/site/simonbarthelme/software`, that implements EP-ABC for several models, including spatial extremes. The current version of the package includes the parallel version described in this paper.

An interesting direction for future work is to integrate current developments on model emulators into EP-ABC. Model emulators are ML algorithms that seek to learn a tractable approximation of the likelihood surface from samples [Wilkinson, 2014]. A variant directly learns an ap-

proximation of the posterior distribution, as in Gutmann and Corander [2015]. Heess et al. [2013] introduce a more direct way of using emulation in an EP context. Their approach is to consider each site as a mapping between the parameters of the pseudo-prior and the mean and covariance of the hybrid, and to learn the parameters of that mapping. In complex but low-dimensional models typical of ABC applications this viewpoint could be very useful and deserves to be further explored.

Acknowledgments

We are very grateful to Mark Beaumont, Dennis Prangle, and Markus Hainy for their careful reading and helpful comments that helped us to improve this chapter.

The second author is partially supported by ANR (Agence Nationale de la Recherche) grant ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047 as part of Investissements d'Avenir program.

Bibliography

P. Alquier, V. Cottet, and G. Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *arXiv preprint arXiv:1702.01402*, 2017.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007. ISSN 0090-5364.

Francis R Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9(Jun):1019–1048, 2008.

Franck Barthe, Olivier Guédon, Shahar Mendelson, and Assaf Naor. A probabilistic approach to the geometry of the l_p^n-ball. *The Annals of Probability*, 33(2):480–513, 2005. ISSN 0091-1798.

Simon Barthelmé and Nicolas Chopin. Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109(505):315–333, 2014.

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006. ISSN 0178-8051.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Large margin classifiers: Convex loss, low noise, and convergence rates. In *NIPS*, pages 1173–1180, 2003.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005. ISSN 0090-5364.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009. ISSN 0006-3444.

A. Belloni and V. Chernozhukov. ℓ-1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103, 2015.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005. ISSN 1292-8100.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1): 1–122, 2011.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

T. Cai and W.-X. Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14: 3619–3647, 2013.

Emmanuel J Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization.* Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004.

O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37 of *Panoramas et Synthèses [Panoramas and Syntheses]*. Société Mathématique de France, Paris, 2012. ISBN 978-2-85629-370-6.

Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012. ISSN 1615-3375.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 02 2015.

V. Cottet and P. Alquier. 1-bit Matrix Completion: PAC-Bayesian Analysis of a Variational Approximation. *ArXiv e-prints*, April 2016.

Botond Cseke and Tom Heskes. Approximate marginals in latent Gaussian models. *J. Mach. Learn. Res.*, 12:417–454, 2011.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002. ISSN 0273-0979.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1): 39–61, 2008. ISSN 1573-0565.

Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

Thomas A. Dean, Sumeetpal S. Singh, Ajay Jasra, and Gareth W. Peters. Parameter estimation for hidden Markov models with intractable likelihoods. *Scand. J. Stat.*, 41(4):970–987, 2014. ISSN 0303-6898.

Guillaume Dehaene and Simon Barthelmé. Expectation propagation in the large-data limit. *arXiv preprint arXiv:1503.08060*, 2015a.

Guillaume P Dehaene and Simon Barthelmé. Bounding errors of expectation-propagation. In *Advances in Neural Information Processing Systems*, pages 244–252, 2015b.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.*, 22(5):1009–1020, 2012.

R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-00754-2. Revised reprint of the 1989 original.

Lutz Dümbgen. Bounding standard gaussian tail probabilities. Technical report, University of Bern, 2010.

Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *arXiv preprint arXiv:1603.09071*, 2016.

Robert J. Erhardt and Richard L. Smith. Approximate bayesian computing for spatial extremes. *Computational Statistics & Data Analysis*, 56(6):1468 – 1481, 2012. ISSN 0167-9473.

Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

Manuel Garcia-Magariños, Anestis Antoniadis, Ricardo Cao, and Wenceslao González-Manteiga. Lasso logistic regression, GSoft and the cyclic coordinate descent algorithm: application to gene expression data. *Stat. Appl. Genet. Mol. Biol.*, 9:Art. 30, 30, 2010. ISSN 1544-6115.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2014. ISBN 978-1-4398-4095-5.

Michael U. Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *ArXiv preprint 1501.03291*, January 2015.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Nicolas Heess, Daniel Tarlow, and John Winn. Learning to pass expectation propagation messages. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3219–3227. Curran Associates, Inc., 2013.

R. Herbrich and T. Graepel. A PAC-Bayesian margin bound for linear classifiers. *IEEE Transactions on Information Theory*, 48(12):3140–3150, 2002.

Mark Herbster, Stephen Pasteris, and Massimiliano Pontil. Mistake bounds for binary matrix completion. In *Advances In Neural Information Processing Systems*, pages 3954–3962, 2016.

Cho-Jui Hsieh and Peder A Olsen. Nuclear norm minimization via active subspace selection. In *ICML*, pages 575–583, 2014.

Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit S Dhillon. PU learning for matrix completion. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2445–2453, 2015.

Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964. ISSN 0003-4851.

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Ajay Jasra. Approximate Bayesian computation for a class of time series models. *International Statistical Review*, 2015.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Olga Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electronic journal of statistics*, 9(2):2348–2369, 2015.

Olga Klopp, Jean Lafond, Éric Moulines, Joseph Salmon, et al. Adaptive multinomial matrix completion. *Electronic Journal of Statistics*, 9(2): 2950–2975, 2015.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 2002. ISSN 0090-5364.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006. ISSN 0090-5364.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. ISBN 978-3-642-22146-0. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329, 2011.

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–412, 2010.

Jean Lafond, Olga Klopp, Eric Moulines, and Joseph Salmon. Probabilistic low-rank matrix completion on finite alphabets. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2014.

Pierre Latouche, Stéphane Robin, and Sarah Ouadah. Goodness of fit of logistic models for random graphs. *arXiv preprint arXiv:1508.00286*, 2015.

Guillaume Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007. ISSN 1350-7265.

Guillaume Lecué. *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis*. Habilitation à Diriger des Recherches Université. Paris-Est Marne-la-vallée, December 2011.

Guillaume Lecué and Shahar Mendelson. General nonexact oracle inequalities for classes with a subexponential envelope. *The Annals of Statistics*, 40(2):832–860, 2012. ISSN 0090-5364.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method II: complexity dependent error rates. Technical report, CNRS, Ecole Polytechnique and Technion, 2015a.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: sparse recovery. Technical report, CNRS, Ecole Polytechnique and Technion, 2015b.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. ISBN 3-540-52013-9. Isoperimetry and processes.

Christiane Lemieux. *Monte Carlo and Quasi-Monte Carlo Sampling (Springer Series in Statistics)*. Springer, feb 2009.

Yew Jin Lim and Yee Whye Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD cup and workshop*, volume 7, pages 15–21, 2007.

The Tien Mai and Pierre Alquier. A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9:823–841, 2015.

Carmen Mak. *Polychotomous logistic regression via the Lasso*. ProQuest LLC, Ann Arbor, MI, 1999. ISBN 978-0612-41227-9. Thesis (Ph.D.)–University of Toronto (Canada).

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

D.A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York, 1998. ACM.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1): 53–71, 2008. ISSN 1369-7412.

Shahar Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002.

Shahar Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4(5):759–771, 2004. ISSN 1532-4435.

Shahar Mendelson. Obtaining fast error rates in nonconvex situations. *J. Complexity*, 24(3):380–397, 2008. ISSN 0885-064X.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Ann. Statist.*, 38(1):526–565, 2010. ISSN 0090-5364.

T. Minka. Power EP. Technical report, Dep. Statistics, Carnegie Mellon University, Pittsburgh, PA., 2004.

T. P. Minka. Expectation propagation for approximate Bayesian inference. *Proceedings of Uncertainty in Artificial Intelligence*, 17:362–369, 2001.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

S. A. Padoan, M. Ribatet, and S. A. Sisson. Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.*, 105(489):263–277, 2010. ISSN 0162-1459.

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Dante J Paz and Ariel G Sánchez. Improving the precision matrix for precision cosmology. *Monthly Notices of the Royal Astronomical Society*, 454(4):4326–4334, 2015.

Dennis Prangle. Lazy ABC. *Statistics and Computing*, pages 1–15, dec 2014.

M. M. Rao and Z. D. Ren. *Theory of Orlicz spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 1991. ISBN 0-8247-8478-2.

M. M. Rao and Z. D. Ren. *Applications of Orlicz spaces*, volume 250 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 2002. ISBN 0-8247-0730-3.

Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387212396.

Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Erlis Ruli, Nicola Sartori, and Laura Ventura. Approximate Bayesian computation with composite score functions. *Statistics and Computing*, pages 1–14, 2015. ISSN 0960-3174.

N. Sabbe, O. Thas, and J.-P. Ottoy. EMLasso: logistic lasso with missing data. *Stat. Med.*, 32(18):3143–3157, 2013. ISSN 0277-6715.

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov Chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887. ACM, 2008.

Martin Schlather. Models for stationary max-stable random fields. *Extremes*, 5(1):33–44, 2002. ISSN 1386-1999.

M. Seeger. Expectation propagation for exponential families. Technical report, Univ. California Berkeley, 2005.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

Yevgeny Seldin and Naftali Tishby. PAC–Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11 (Dec):3595–3646, 2010.

J. Shawe-Taylor and J. Langford. PAC-Bayes & margins. *Advances in neural information processing systems*, 15:439, 2003.

S. A. Sisson, Y. Fan, and Mark M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104(6):1760–1765 (electronic), 2007. ISSN 1091-6490.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2005.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008. ISBN 978-0-387-77241-7.

Weijie Su and Emmanuel Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016. ISSN 0090-5364.

Michel Talagrand. *The generic chaining.* Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. ISBN 3-540-24518-9. Upper and lower bounds of stochastic processes.

Guo-Liang Tian, Man-Lai Tang, Hong-Bin Fang, and Ming Tan. Efficient methods for estimating constrained parameters with applications to regularized (lasso) logistic regression. *Comput. Statist. Data Anal.*, 52 (7):3528–3542, 2008. ISSN 0167-9473.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. ISSN 0090-5364.

Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics, 2009.

Sara van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, [Cham], 2016. ISBN 978-3-319-32773-0; 978-3-319-32774-7. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000. ISBN 0-521-65002-X.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

Vladimir N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. ISBN 0-471-03003-1. A Wiley-Interscience Publication.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statist. Sinica*, 21(1):5–42, 2011. ISSN 1017-0405.

George Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992.

S. R. White, T. Kypraios, and S. P. Preston. Piecewise Approximate Bayesian Computation: fast inference for discretely observed Markov models using a factorised posterior distribution. *Stat. Comput.*, 25(2): 289–301, 2015. ISSN 0960-3174.

Richard D. Wilkinson. Accelerating ABC methods using Gaussian processes. *ArXiv preprint 1401.1436*, February 2014.

Sinan Yıldırım, Sumeetpal S Singh, Thomas Dean, and Ajay Jasra. Parameter estimation in hidden Markov models with intractable likelihoods using sequential Monte Carlo. *Journal of Computational and Graphical Statistics*, 24(3):846–865, 2014.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. ISSN 1467-9868.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32 (1):56–85, 02 2004.

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

Titre : Étude théorique de quelques procédures statistiques pour le traitement de données complexes

Mots Clefs : Statistiques, Inférence Bayésienne, Statistiques computationnelles, Machine Learning, Complétion de matrices, Extrêmes spatiaux

Résumé : La partie principale de cette thèse s'intéresse à développer les aspects théoriques et algorithmiques pour trois procédures statistiques distinctes. Le premier problème abordé est la complétion de matrices binaires. Nous proposons un estimateur basé sur une approximation variationnelle pseudo-bayésienne en utilisant une fonction de perte différente de celles utilisées auparavant. Nous pouvons calculer des bornes non asymptotiques sur le risque intégré. L'estimateur proposé est beaucoup plus rapide à calculer qu'une estimation de type MCMC et nous montrons sur des exemples qu'il est efficace en pratique. Le deuxième problème abordé est l'étude des propriétés théoriques du minimiseur du risque empirique pénalisé pour des fonctions de perte lipschitziennes. Nous pouvons ensuite appliquer les résultats principaux sur la régression logistique avec la pénalisation SLOPE ainsi que sur la complétion de matrice. Le troisième chapitre développe une approximation de type Expectation-Propagation quand la vraisemblance n'est pas explicite. On utilise alors l'approximation ABC dans un second temps. Cette procédure peut s'appliquer à beaucoup de modèles et est beaucoup plus précise et rapide. Elle est appliquée à titre d'exemple sur un modèle d'extrêmes spatiaux.

Title : Theoretical study of some statistical procedures applied to complex data

Keys words : Statistics, Bayesian Inference, Computational Statistics, Machine Learning, Matrix Completion, Spatial Extremes

Abstract :
The main part of this thesis aims at studying the theoretical and algorithmic aspects of three distinct statistical procedures. The first problem is the binary matrix completion. We propose an estimator based on a variational approximation of a pseudo-Bayesian estimator. We use a different loss function of the ones used in the literature. We are able to compute non asymptotic risk bounds. It is much faster to compute the estimator than a MCMC method and we show on examples that it is efficient in practice. In a second part we study the theoretical properties of the regularized empirical risk minimizer for Lipschitz loss functions. We are therefore able to apply it on the logistic regression with the SLOPE regularization and on the matrix completion as well. The third chapter develops an Expectation-Propagation approximation when the likelihood is not explicit. We then use an ABC approximation in a second stage. This procedure may be applied to many models and is more precise and faster than the classic ABC approximation. It is used in a spatial extremes model.