Complétion de matrice : aspects statistiques et computationnels

Jean Lafond

NNT : 2016SACLT002

Thèse de doctorat de l'Université Paris-Saclay préparée Télecom ParisTech

Ecole doctorale n°574
Mathématiques Hadamard
Spécialité de doctorat : Mathématiques appliquées

par

M. Jean Lafond

Complétion de Matrice de Faible Rang: Aspects Statistiques et Computationnels

Thèse présentée et soutenue à Télecom ParisTech, le 19 Décembree 2016.

Composition du Jury :

M.	Stéphan Clémençon	Professeur Télécom ParisTech	(Examinateur)
M.	Anatoli Juditsky	Directeur de recherche INRIA	(Rapporteur)
Mme.	Olga Klopp	Maître de Conférence Université Paris Ouest	(Examinatrice)
M.	Eric Moulines	Professeur Polytechnique	(Directeur de thèse)
M.	Vincent Rivoirard	Professeur Université Paris Dauphine	(Rapporteur)
M.	Joseph Salmon	Maître de Conférence Télécom ParisTech	(Directeur de thèse)

NNT : 2016SACLT002

Thèse de doctorat de l'Université Paris-Saclay préparée Télecom ParisTech

Ecole doctorale n°574
Mathématiques Hadamard
Spécialité de doctorat : Mathématiques appliquées

par

M. Jean Lafond

Low Rank Matrix Completion: Statistical and Computational Aspects

Thèse présentée et soutenue à Télecom ParisTech, le 19 Décembree 2016.

Composition du Jury :

M.	Stéphan Clémençon	Professeur Télécom ParisTech	(Examinateur)
M.	Anatoli Juditsky	Directeur de recherche INRIA	(Rapporteur)
Mme.	Olga Klopp	Maître de Conférence Université Paris Ouest	(Examinatrice)
M.	Eric Moulines	Professeur Polytechnique	(Directeur de thèse)
M.	Vincent Rivoirard	Professeur Université Paris Dauphine	(Rapporteur)
M.	Joseph Salmon	Maître de Conférence Télécom ParisTech	(Directeur de thèse)

A la mémoire de Ghislaine Lafond,
Anne et Jean Guichebaron

Remerciements

Je tiens à remercier en premier lieu mes deux directeurs de thèse Eric et Joseph. Merci Eric d'avoir pris le temps de partager une partie de ton immense culture scientifique. Merci de m'avoir orienté vers des sujets intéressants et de m'avoir appris le métier de chercheur. Merci aussi pour toutes les conversations que nous avons pu avoir, sur la science ou la vie en générale. Merci Joseph pour le temps que tu as passé avec moi à écrire des équations sur l'immense tableau qui siège dans ton bureau. Merci pour les (très) nombreuses relectures et corrections que tu as faites pendant ma thèse. Merci pour les heures passées à la terrasse du Viking dans la fraîcheur cantalienne.

Je tiens à remercier Anatoli Juditsky et Vincent Rivoirard pour avoir accepté de rapporter ma thèse. Merci pour tous vos commentaires et remarques qui m'ont permis d'améliorer le manuscrit.

Merci aux examinateurs Olga Klopp et Stephan Clémençon qui m'ont fait l'honneur d'assister à ma soutenance de thèse. Merci Olga de m'avoir introduit au sujet de complétion de matrices et pour le travail que nous avons fait ensemble.

Merci à Léon Bottou de m'avoir encadré pendant mon stage de recherche réalisé en fin de thèse. Merci de m'avoir transmis tant de connaissances sur l'optimisation et les réseaux de neurones.

Merci aux membres du laboratoire LTCI avec qui j'ai vécu pendant trois ans et partagé tant de repas dans l'inénarrable FIAP. Merci à Minh et Guillaume qui m'ont supporté comme co-bureau.

Merci à mes amis de lycée, de prépas et d'ailleurs. Merci pour les moments passés ensemble qui m'ont permis de m'évader de ma thèse.

Merci à mes anciens professeurs. En particulier, je tiens à remercier mes professeurs de mathématiques de prépas, messieurs Testud et Abou-Jaoudé.

Merci à mes parents, à mes sœurs, à mon grand père, à mon parrain, à ma famille pour tout ce qu'il m'ont apportés depuis de si nombreuses années.

Enfin, merci à toi Caroline qui m'a soutenu depuis le début lorsque j'ai eu cette idée un peu folle de vouloir faire une thèse. Merci d'avoir supporté mes nombreux déplacements loin de la maison. En fait, merci pour tout.

Contents

List of Figures ix

List of Tables xi

1 Etat de l'art et contributions 1
- 1.1 Contexte et état de l'art . . . 2
  - 1.1.1 Un exemple introductif . . . 2
  - 1.1.2 Les méthodes de complétion sous hypothèse de rang faible . . . 4
  - 1.1.3 Garanties de reconstruction . . . 6
  - 1.1.4 Méthodes d'optimisation pour la complétion de matrice . . . 10
- 1.2 Contributions . . . 15
  - 1.2.1 Complétion de matrice pour les modèles multinomiaux . . . 15
  - 1.2.2 Complétion de matrice avec bruit appartenant à la famille exponentielle 17
  - 1.2.3 Algorithmes de Frank Wolfe stochastiques . . . 19

2 Adaptive multinomial matrix completion 23
- 2.1 Introduction . . . 24
- 2.2 Main results . . . 26
  - 2.2.1 One-bit matrix completion . . . 26
  - 2.2.2 Minimax lower bounds for one-bit matrix completion . . . 28
  - 2.2.3 Extension to multi-class problems . . . 29
- 2.3 Implementation . . . 30
- 2.4 Numerical Experiments . . . 32
- 2.5 Proofs of main results . . . 34
  - 2.5.1 Proof of Theorem 2.2 and Theorem 2.6 . . . 34
  - 2.5.2 Proof of Theorem 2.7 . . . 40
  - 2.5.3 Proof of Theorem 2.5 . . . 43

3 Low rank matrix completion with exponential family noise 47
- 3.1 Introduction . . . 48
- 3.2 Main results . . . 50
  - 3.2.1 Model Specification . . . 50
  - 3.2.2 General Matrix Completion . . . 50
  - 3.2.3 Matrix Completion with known sampling scheme . . . 53
  - 3.2.4 Lower Bound . . . 54
- 3.3 Proofs of main results . . . 55
  - 3.3.1 Proof of Theorem 3.5 . . . 55
  - 3.3.2 Proof of Theorem 3.6 . . . 57
  - 3.3.3 Proof of Theorem 3.14 . . . 58

3.4 Proof of Theorem 3.16 and Theorem 3.17 59
3.5 Proof of Theorem 3.19 60
3.6 Proof of Oracle inequalities and Bounds for Completion with known sampling 61
3.6.1 Proof of Theorem 3.9 61
3.6.2 proof of Theorem 3.11 63

4 On the Stochastic Frank-Wolfe Algorithms for Convex and Non-convex Optimization 65
4.1 Introduction 66
4.2 Problem Setup and Algorithms 68
4.3 Convergence Analysis of S-FW and S-AW 71
4.3.1 Convex Optimization 71
4.3.2 Non-convex Optimization 73
4.4 Application: Online Learning 74
4.4.1 Convex Loss 75
4.4.2 Non-convex Loss 76
4.5 Numerical Experiments 77
4.5.1 Example: Online LASSO 77
4.5.2 Example: Online matrix completion (MC) 79
4.5.3 Example: Robust Binary Classification with Outliers 81
4.6 Proofs 82
4.6.1 Proof of Theorem 3 82
4.6.2 Proof of Theorem 4 84
4.6.3 Proof of Theorem 5 91
4.6.4 Proof of Proposition 6 95
4.6.5 Proof of Proposition 10 96
4.6.6 Proof of Proposition 11 97
4.6.7 Fast convergence of S-AW without strong convexity 99

Bibliography 101

List of Figures

1.1 Exemple de matrice et schéma d'observation (ligne, colonne, note) 3
1.2 Mise à jour de l'itéré dans l'algorithme de Frank Wolfe 13

2.1 Kullback-Leibler divergence between the estimated and the true model for different matrices sizes and sampling fraction, normalized by number of classes. Right figure: binomial and the Gaussian models ; left figure: multinomial with five classes and Gaussian model. 34

4.1 Online LASSO with synthetic data. Convergence of the primal optimality for online LASSO with (Left) $r = 1.1\|\bar{\boldsymbol{\theta}}\|_1 > \|\boldsymbol{\theta}^\star\|_1$; (Right) $r = 0.15\|\bar{\boldsymbol{\theta}}\|_1 = \|\boldsymbol{\theta}^\star\|_1$. 78
4.2 Online LASSO with single-pixel imaging data `R64.mat`. (Left) Convergence of the objective value. (Middle) Reconstructed image after 500 iterations of O-FW; (Right) O-AW. 78
4.3 Online MC performance. (Left) synthetic with batch size $B = 1000$; (Middle) `movielens100k` with $B = 80$; (Right) `movielens20m` with $B = 10000$. (Top) objective value/MSE against round number; (Bottom) against execution time. The duality gap g_t^{FW} is plotted in purple. 80
4.4 Binary classification performance against round number t for: (Left) synthetic data; (Middle) `mnist` (class '1'); (Right) `rcv1.binary`. (Top) with no flip (Bottom) with 25% flip in the training labels. The duality gap g_t^{FW} for O-FW with sigmoid loss is plotted in purple. 81

List of Tables

1.1 Notations 2
1.2 Dimensions de quelques jeux de données représentatifs en recommandation . 3
1.3 Execution time of the proposed algorithm for the binary case. 17
1.4 Quelques distributions de la famille exponentielle 18
1.5 Vitesse de convergence donnée dans le contexte de l'optimisation stochastique et en ligne (regret) 22

2.1 Execution time of the proposed algorithm for the binary case. 32
2.2 Prediction errors for a binomial (2 classes) underlying model, for a 1000×600 matrix. 33
2.3 Prediction Error for a multinomial (5 classes) distribution against a 1000×600 matrix. 34
2.4 Binomial prediction error when performing one versus the others procedure on the MovieLens $100k$ dataset. 34

3.1 Parametrization of some exponential family distributions 50

4.1 Convergence rate comparison. Note that the regret bound for Garber & Hazan (2015b) is given under an adversarial loss setting, while the bounds for Hazan & Kale (2012) and our work are based on a stochastic cost. Depending on the applications (see Section 4.5 & subsection 4.5.1), our regret and anytime bounds can be improved to $\mathcal{O}(\log^2 t/t)$ and $\mathcal{O}(\log t/t)$, respectively. 67

CHAPTER 1
Etat de l'art et contributions

1.1 Contexte et état de l'art

Notations Nous résumons dans le tableau ci-dessous les notations utilisées dans ce chapitre.

Notation	Description
$\mathbb{R}$	ensembles des réels
$\mathbb{N}$	ensemble des entiers naturels
$[n]$	ensemble $\{1, \dots, n\}$ avec $n \in \mathbb{N}$
$\mathbb{R}^{m_1 \times m_2}$	ensemble des matrices à m_1 lignes m_2 colonnes
$\mathbb{R}^{m_1 \times m_2 \times q}$	$(\mathbb{R}^{m_1 \times m_2})^q$
$X_{k\cdot}$ (*resp.* $X_{\cdot l}$)	ligne k (*resp.* colonne l) d'une matrice X
$X^\top$	transposée d'une matrice X
$\mathrm{trace}(\cdot)$	trace d'une matrice
$\mathrm{rk}(\cdot)$	rang d'une matrice
$\mathrm{diag}(u)$	matrice diagonale dont la diagonale est le vecteur u
$\|\cdot\|_{\sigma,p}$	Schatten p-norme de matrices
$\|\cdot\|_{\sigma,\infty}$	norme opérateur de matrice
$\|\cdot\|_\infty$	norme infinie d'une matrice $\|X\|_\infty = \max_{i,j} \|X_{i,j}\|$
$\langle \cdot, \cdot \rangle$	produit scalaire, pour les matrices $\langle X, Y \rangle := \mathrm{trace}(XY^\top)$
$\mathbb{P}$	probabilité
$\mathbb{E}$	espérance
$\nabla \cdot$	gradient
$\mathcal{O}(\cdot)$	domination

TABLE 1.1: Notations

1.1.1 Un exemple introductif

Beaucoup de problèmes d'apprentissage peuvent se reformuler comme un problème de complétion de matrice. Celui de la recommandation de film est sans doute le plus célèbre d'entre eux car il a été popularisé par Netflix (un loueur de films en ligne) qui lança en 2006 un concours afin d'améliorer la prédiction des goûts cinématographiques de ses utilisateurs (voir Bell & Koren (2007)). Pour ce faire, les concurrents disposaient (entre autres) de l'historique des notes attribuées par certains utilisateurs à certains films. Évidemment, ces observations sont très parcellaires puisque le nombre de films notés par un utilisateur est bien inférieur au nombre total de films. On peut donc se poser la question suivante: **est il possible de prévoir la note qu'un utilisateur donnerait à un film qu'il n'a pas encore noté?**

Cette question se prête naturellement à la modélisation matricielle. En effet, notons m_1 le nombre d'utilisateurs et m_2 le nombre de films. On peut représenter les notes données par les utilisateurs aux films par une matrice $X \in \mathbb{R}^{m_1 \times m_2}$, où une entrée $X_{k,l}$ représente la note donnée par l'utilisateur k au film l. Une observation est alors donnée par un indice de ligne (*i.e.,* un utilisateur), un indice de colonne (*i.e.,* un film) et la valeur de l'entrée X correspondante (*i.e.,* la note), cf Figure 1.1.

Avec ce formalisme, la question de prédiction devient un cas particulier de la question plus générale suivante:

Problème 1 (Problème de complétion de matrice)**.** *Comment peut-on reconstruire une matrice à partir d'une observation partielle et possiblement bruitée de ses entrées?*

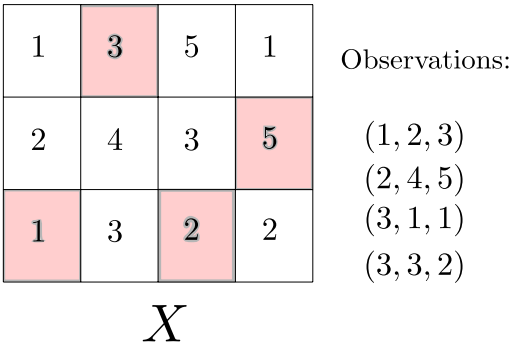

FIGURE 1.1: Exemple de matrice et schéma d'observation (ligne, colonne, note)

C'est précisément à cette question que les méthodes de complétion de matrice tentent de répondre. Néanmoins, afin d'espérer résoudre les problèmes de recommandation de films (entre autres), il est crucial de prendre en compte certaines contraintes.

D'une part, les problèmes de complétion de matrices sont souvent mal posés. En effet, si l'on note n le nombre d'observations, le Tableau 1.2 montre que $n \ll m_1 m_2$ pour les bases de données de recommandation de films. Autrement dit, le nombre d'observations est bien inférieur au nombre total d'entrées à reconstruire et par conséquent il est nécessaire de réduire la dimension du problème pour espérer le résoudre. Dans ce mémoire, nous nous intéresserons au cas où la matrice X est de **faible rang**. Comme on le verra dans la suite de ce mémoire, il est possible de tirer parti de cette hypothèse pour reconstruire la matrice inconnue, même lorsque le nombre d'observations est faible devant le nombre total de coefficients de la matrice.

D'autre part, les dimensions des problèmes traités en pratique peuvent être grandes. Ainsi, quelle que soit la méthode de complétion choisie, il est crucial qu'elle puisse **passer à l'échelle**.

Jeu de données	m_1	m_2	n
MovieLens 20M	$138 \cdot 10^3$	$27 \cdot 10^3$	$20 \cdot 10^6$
NetFlix	$780 \cdot 10^3$	$17 \cdot 10^3$	$100 \cdot 10^6$
Yahoo! Music	$1.8 \cdot 10^6$	$137 \cdot 10^3$	$717 \cdot 10^6$

TABLE 1.2: Dimensions de quelques jeux de données représentatifs en recommandation

Cet exemple introductif permet de mettre en lumière deux aspects importants des problèmes de complétion de matrice. D'abord un aspect **statistique**, qui renvoie à la qualité de reconstruction de la matrice inconnue $\bar{X}$; ensuite un aspect **numérique** qui interroge la capacité des méthodes de complétion à être implémentées sur de grands jeux de données. Nous traiterons ces deux aspects dans ce mémoire.

Nous renvoyons à Koren et al. (2009); Bobadilla et al. (2013) pour une présentation détaillée des applications des méthodes de complétion au systèmes de recommandation. Enfin, notons que les méthodes de complétion de matrices de faible rang s'appliquent à bien d'autres problèmes. Parmi les autres applications, on peut citer entre autres: la tomographie quantique (*e.g.,* Gross (2011)), la reconstruction d'images (*e.g.,* Hui et al. (2010); Ji et al. (2013); Xu et al. (2014)), la reconstruction de données sismiques (*e.g.,* Yang et al. (2013)).

1.1.2 Les méthodes de complétion sous hypothèse de rang faible

On considère une matrice $\bar{X} \in \mathbb{R}^{m_1 \times m_2}$ et un échantillon de n observations de la forme $(k_i, l_i, Y_i)_{i=1}^n$ avec $k_i \in [m_1]$, $l_i \in [m_2]$ et $Y_i \in \mathbb{R}$. Chaque valeur Y_i correspond à une observation bruitée du coefficient $\bar{X}_{k_i,l_i}$. Dans un premier temps, nous supposerons qu'il s'agit d'un bruit additif (des modèles plus généraux seront présentés par la suite) *i.e.,*

$$Y_i = \bar{X}_{k_i,l_i} + \epsilon_i \ , \tag{1.1}$$

avec ϵ_i un bruit centré ($\mathbb{E}[\epsilon_i] = 0$). Le but des méthodes de complétion est de produire un estimateur $\hat{X}$ de $\bar{X}$ à partir des observations $(k_i, l_i, Y_i)_{i=1}^n$.

Cependant, comme discuté précédemment, les problèmes de complétion de matrice sont en général mal posés puisque le nombre d'observations vérifie $n \ll m_1 m_2$. Afin de réduire la dimension du problème, on peut supposer que la matrice $\bar{X}$ est de faible rang. En effet, si l'on note r le rang de $\bar{X}$, alors une décomposition en valeur singulière montre que l'on a la factorisation suivante:

$$\bar{X} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top \ ,$$

avec $(\sigma_i)_{i=1}^r$ les valeurs singulières de $\bar{X}$ et $(u_i, v_i)_{i=1}^r$ ses vecteurs singuliers. Avec cette forme factorisée il est clair que le nombre de degrés de libertés passe de $m_1 m_2$ à $\mathcal{O}(r(m_1 + m_2))$. Ainsi, lorsque le rang r est suffisamment faible pour que l'on ait $n \geq r(m_1 + m_2)$, on peut espérer construire un bon estimateur de $\bar{X}$. Dans cette section, nous présentons différents estimateurs largement utilisés en pratique pour reconstruire $\bar{X}$.

Modèle factorisé

Une manière d'estimer $\bar{X}$ consiste à fixer une borne supérieure R sur le rang de $\bar{X}$ et ensuite à chercher la meilleure approximation de rang au plus R, qui minimise l'attache quadratique aux données:

$$\hat{X} := \underset{X \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \ \sum_{i=1}^{n} (Y_i - X_{k_i,l_i})^2 \quad \text{s.c.} \quad \mathrm{rk}(X) \leq R \ . \tag{1.2}$$

Le problème (1.2) est malheureusement connu pour être NP-difficile (voir Candès & Recht (2009)). Cependant, on peut le reformuler en cherchant deux facteurs $U \in \mathbb{R}^{m_1 \times R}$ et $V \in \mathbb{R}^{m_2 \times R}$ minimisant

$$\hat{X} := \hat{U}\hat{V}^\top \quad \text{avec} \quad \hat{U}, \hat{V} := \underset{\substack{U \in \mathbb{R}^{m_1 \times R} \\ V \in \mathbb{R}^{m_2 \times R}}}{\arg\min} \ \sum_{i=1}^{n} \left(Y_i - U V_{k_i,l_i}^\top \right)^2 \ . \tag{1.3}$$

Le problème (1.3) est biconvexe mais n'est pas convexe en U et en V, néanmoins il est différentiable et l'on peut en trouver un minimum local de façon efficace (voir la section 1.1.4).

Le principal inconvénient de cette méthode est qu'elle nécessite une connaissance sur le rang de la matrice inconnue $\bar{X}$ qui, *a priori*, n'est pas disponible. Cependant, sa facilité d'implémentation la rend très utilisée en pratique (voir Koren et al. (2009)).

Modèle à pénalité norme trace

Une autre approche consiste à considérer une relaxation convexe au problème (1.2) basée sur la norme trace (définie comme la somme des valeurs singulières d'une matrice). On s'intéresse ainsi au problème suivant:

$$\hat{X} := \underset{X \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \sum_{i=1}^{n} (Y_i - X_{k_i,l_i})^2 + \lambda \|X\|_{\sigma,1} \ , \tag{1.4}$$

avec $\lambda > 0$. Un parallèle peut être fait avec la régression LASSO (Tibshirani (1996)). En effet, de même que la norme ℓ_1 est utilisée pour recouvrir des solutions parcimonieuses dans les problèmes de régression, la norme trace (qui n'est rien d'autre que la norme ℓ_1 des valeurs singulières d'une matrice) sert à reconstruire des matrice de rang faible dans les problèmes de complétion.

Le paramètre λ joue un rôle crucial puisqu'il vient équilibrer l'importance donnée à l'attache aux données (terme de perte quadratique) par rapport à la pénalité norme trace. Ainsi, lorsque le bruit d'observation est faible (*i.e.,* $|\epsilon_i| \ll 1$) il est naturel de considérer le problème de reconstruction exact suivant qui est un cas limite du précédent (pour $\lambda \to 0$):

$$\hat{X} := \underset{X \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \ \|X\|_{\sigma,1} \quad \text{s.c.} \quad Y_i = X_{k_i,l_i} \ \forall i \in [n] \ . \tag{1.5}$$

Contrairement au problème (1.2), les problèmes (1.4) et (1.5) sont convexes. En outre, ils admettent aussi une reformulation factorisée. En effet, on peut montrer que la norme trace vérifie l'inégalité suivante (voir Recht et al. (2010)):

$$\|X\|_{\sigma,1} = \min_{\substack{U \in \mathbb{R}^{m_1 \times R} \\ V \in \mathbb{R}^{m_2 \times R}}} \frac{1}{2}(\|U\|_{\sigma,2}^2 + \|V\|_{\sigma,2}^2) \quad \text{s.c.} \quad X = UV^\top \ ,$$

et donc reformuler le problème (1.4) en

$$\hat{U}, \hat{V} := \underset{\substack{U \in \mathbb{R}^{m_1 \times R} \\ V \in \mathbb{R}^{m_2 \times R}}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - UV^\top_{k_i,l_i}\right)^2 + \frac{\lambda}{2}\left(\|U\|_{\sigma,2}^2 + \|V\|_{\sigma,2}^2\right) \ . \tag{1.6}$$

L'intérêt de cette reformulation n'apparaît pas immédiatement puisqu'elle est non convexe. Néanmoins, comme nous le verrons en Section 1.1.4, elle permet l'implémentation d'algorithmes d'optimisation efficaces pour les gros jeux de données.

Modèle de bruit général

Les problèmes de complétion de matrice concernent un grand nombre d'applications avec différents types de données et distributions: catégorielle, comptage, continue, etc.. A cet égard, la structure de bruit additif centré (1.1) peut apparaître restrictive.

Elle peut être généralisée en supposant que la valeur Y_i suit une distribution paramétrée par $\bar{X}_{k_i,l_i}$:

$$Y_i | k_i, l_i \sim \mathbb{P}_{\bar{X}_{k_i,l_i}} \ . \tag{1.7}$$

Cette formulation inclut bien évidemment le cas du bruit additif. Une application importante introduite par Davenport et al. (2014) est celle des méthodes de complétion pour les données

binaires où $Y_i \in \{0, 1\}$. Dans ce cas là, on peut supposer que l'on a $\mathbb{P}(Y_i|k_i, l_i) = f(\bar{X}_{k_i,l_i})$ avec f une fonction de lien (*e.g.,* logistique ou probit). Il est important de noter que pour les modèles généraux, l'espace auquel appartient le paramètre $\bar{X}$ n'est plus nécessairement le même que celui des observations Y_i.

Afin de prendre en compte la distribution considérée, la fonction d'attache aux données n'est plus nécessairement la perte quadratique. Elle est remplacée par une fonction L qui est plus adaptée (*e.g.,* la log vraisemblance négative de la distribution):

$$\hat{X} := \underset{X \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \sum_{i=1}^{n} L(Y_i, X_{k_i,l_i}) + \lambda \|X\|_{\sigma,1} \, . \tag{1.8}$$

1.1.3 Garanties de reconstruction

Dans cette section, nous nous intéressons aux garanties statistiques qui existent sur l'erreur de reconstruction de la matrice $\bar{X}$ par l'estimateur $\hat{X}$. Pour les méthodes de reconstruction inexactes (*i.e.,* lorsque l'on ne cherche pas à reconstruire exactement la matrice $\bar{X}$), nous chercherons à borner avec grande probabilité le risque quadratique:

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \, .$$

Nous concluons cette section en introduisant les différentes problématiques associées que nous avons choisi d'étudier dans cette thèse.

Reconstruction exacte

Si les estimateurs pénalisés par la norme trace sont utilisés depuis un certains temps pour les problèmes de complétion de matrice de rang faible (voir par exemple Fazel (2002); Srebro (2004)), les résultats théoriques concernant leur performance sont beaucoup plus récents. Les premiers travaux se sont intéressés au cas de la complétion exacte (sans bruit d'observation) et ont analysé l'estimateur (1.5). Ils ont été initiés par Recht et al. (2010) puis améliorés par Candès & Recht (2009); Candès & Tao (2010). Les preuves ont été significativement simplifiées par Gross (2011) puis Recht (2011) en supposant l'échantillonnage des entrées $(k_i, l_i)_{i=1}^n$ i.i.d. et donc avec la possibilité d'observer deux fois la même entrée. Nous présentons ici les résultats de Recht (2011).

Avant de présenter les résultats, nous devons définir la notion de cohérence d'une matrice. Cette notion est importante car elle permet de caractériser les matrices qui peuvent être reconstruites. En effet, il est clair que certaines matrices ne peuvent pas être reconstruites à moins d'échantillonner toutes leurs entrées. Par exemple, la matrice de rang 1 ayant toutes ses entrées nulles sauf une valant 1 ne pourra être reconstruite proprement que lorsque son entrée non nulle est observée. Intuitivement, le problème avec la matrice précédente est que révéler une entrée ne donne pas d'information sur les autres entrées. C'est cette idée que la cohérence tente de capturer:

Définition 1 (Cohérence)**.** *Etant donné une famille de R vecteurs orthonormée $U \in \mathbb{R}^N$, on définit sa cohérence comme:*

$$\mu(U) = \frac{N}{R} \max_{i=1 \cdots N} \|e_i^\top \Pi_U(e_i)\|_2^2$$

où $(e_i)_{i=1}^N$ *désigne la base canonique de* $\mathbb{R}^N$ *et* Π_U *la projection sur le sous espace vectoriel engendré par* U.

On peut vérifier que $\mu(U) \leq N/R$ et que l'égalité est atteinte s'il existe $i \in [N]$ tel que $e_i \in U$. On a aussi $\mu(U) \geq 1$ avec l'égalité si les coordonnées des vecteurs de U ont même amplitude. Ainsi, pour une famille faiblement cohérente, une projection renseigne sur les autres projections. Par conséquent on peut s'attendre à pouvoir reconstruire d'autant mieux une matrice que ses vecteurs singuliers de droite et de gauche sont faiblement cohérents.

Avant d'énoncer les résultats concernant les garanties de reconstruction, nous aurons besoin des hypothèses suivantes:

H1. *Le modèle d'observation est exact et l'échantillonnage des entrées est i.i.d. et uniforme* i.e.,

$$Y_i = \bar{X}_{k_i,l_i} \quad \textit{avec} \quad (k_i, l_i) \sim \mathcal{U}([m_1] \times [m_2]) \;, \quad \textit{pour} \quad i \in [n] \;.$$

Notons U (*resp.* V) les vecteurs singuliers de gauche (*resp.* de droite) de $\bar{X}$.

H2. *La cohérence de la matrice* $\bar{X}$ *est majorée par* μ_0 i.e., $\max(\mu(U), \mu(V)) \leq \mu_0$

H3. *Il existe une constante* γ *telle que* $\|UV^\top\|_{\sigma_\infty} \leq \gamma\sqrt{\mathrm{rk}(\bar{X})/(m_1 m_2)}$.

On peut alors montrer que sous ces hypothèses, on peut reconstruire $\bar{X}$ avec grande probabilité:

Théorème 1 (Recht (2011))**.** *Soit* $\hat{X}$ *une solution du problème* (1.5)*. Supposons sans perte de généralité que* $m_1 \leq m_2$*. Supposons en outre H1, H2, H3 et*

$$n \geq 32 \max\{\gamma^2, \mu_0\} \,\mathrm{rk}(\bar{X})(m_1 + m_2)\beta \log^2(2m_2) \;,$$

pour un certain $\beta > 1$*. Alors* $\hat{X}$ *est unique et vérifie* $\hat{X} = \bar{X}$ *avec probabilité* $1 - 6\log(m_2)(m_1 + m_2)^{2-2\beta} - m_2^{2-2\beta^{1/2}}$.

Reconstruction inexacte pour modèle à bruit additif

Les garanties statistiques pour les méthodes de reconstruction inexactes ont été étudiées entre autres par Candès & Plan (2010); Keshavan et al. (2010); Gaïffas & Lecué (2011); Koltchinskii et al. (2011); Negahban & Wainwright (2012); Cai & Zhou (2013a); Klopp (2014). Comme son nom l'indique, la reconstruction inexacte n'impose plus à l'estimateur de reconstruire exactement les observation et donc la contrainte $\hat{X}_{k_i,l_i} = Y_i$ disparaît. Nous détaillons ici les résultats obtenus par Klopp (2014).

Nous considérons l'estimateur à pénalité norme trace de la forme suivante:

$$\hat{X} := \underset{\substack{X \in \mathbb{R}^{m_1 \times m_2} \\ \|X\|_\infty \leq \gamma}}{\arg\min} \; \sum_{i=1}^{n} (Y_i - X_{k_i,l_i})^2 + \lambda \|X\|_{\sigma,1} \;. \tag{1.9}$$

Il s'agit de l'estimateur (1.4) avec une contrainte supplémentaire sur l'amplitude maximale des entrées majorée par un paramètre $\gamma > 0$. Même si cette contrainte est importante pour obtenir des garanties théoriques, en pratique on observe qu'elle a peu d'influence et c'est souvent l'estimateur (1.4) qui est utilisé car il est plus simple à calculer.

On suppose que les observations sont générées par un modèle additif de la forme (1.1). La première hypothèse que l'on fait stipule que l'échantillonnage n'est pas trop éloigné d'un échantillonnage uniforme. Notons R_k (*resp.* C_l) la probabilité d'échantillonner une entrée dans la ligne k (*resp.* colonne l):

$$R_k = \sum_{l'=1}^{m_2} \mathbb{P}\left((k_1, l_1) = (k, l')\right) \quad \textit{resp.} \quad C_l = \sum_{k'=1}^{m_1} \mathbb{P}\left((k_1, l_1) = (k', l)\right) ,$$

alors on supposera que l'échantillonnage vérifie l'hypothèse suivante:

H4. *L'échantillonnage $(k_i, l_i)_{i=1}^n$ est i.i.d. En outre il existe μ, $\nu \geq 1$ tels que pour tout $m_1,\ m_2 > 0$ et $(k, l) \in [m_1] \times [m_2]$:*

$$\mathbb{P}\left((k_1, l_1) = (k, l)\right) \geq \frac{1}{\mu m_1 m_2} \quad \textit{et} \quad \max_{k,l}(R_k, C_l) \leq \frac{\nu}{\min(m_1, m_2)}$$

Par ailleurs, on fait aussi les hypothèses suivantes sur le bruit:

H5. *Les bruits $\epsilon_1, \cdots, \epsilon_n$ sont indépendants, centrés ($\mathbb{E}[\epsilon_i] = 0$), de variance $\sigma^2 := \mathbb{E}[\epsilon_i^2]$ et sous exponentiels* i.e., *il existe $K > 0$ tel que*

$$\max_{i \in [n]} \mathbb{E}[\exp(|\epsilon_i|/K)] \leq \mathrm{e} .$$

Sous ces hypothèses on peut alors montrer le résultat suivant.

Théorème 2 (Klopp (2014))**.** *Supposons sans perte de généralité que $m_1 \leq m_2$. Supposons en outre H4, H5 et $\|\bar{X}\|_\infty \leq \gamma$. Soit $\hat{X}$ une solution du problème (1.9) avec*

$$\lambda = 3C^* \sigma \sqrt{\frac{2\nu \log(m_1 + m_2)}{m_1}} .$$

avec C^* une constante dépendant uniquement de K, où K est défini dans H5. Alors, avec une probabilité d'au moins $1 - 3(m_1 + m_2)^{-1}$ on a:

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq c \max\left\{ \max(\gamma^2, \sigma^2)\mu^2 \frac{\mathrm{rk}(\bar{X}) m_2 \log(m_1 + m_2)}{n},\ \gamma^2 \mu \sqrt{\frac{\log(m_1 + m_2)}{n}} \right\} ,$$

avec c une constante numérique absolue.

Reconstruction pour modèle généralisé

Afin de traiter d'autre types de données que celles considérées jusqu'à présent, les estimateurs de la forme (1.8) où la fonction d'attache aux données n'est plus nécessairement quadratique ont été introduits. Le cas de données binaires où la fonction de perte est logistique a été considéré par Davenport et al. (2014), Cai & Zhou (2013b) considèrent aussi une perte logistique mais avec une autre pénalité que la norme trace). Le cas d'observations distribuées selon une loi de la famille exponentielle a été étudié par Gunasekar et al. (2014). Nous reviendrons en détail sur ces modèles aux Chapitres 2 et 3. Nous donnons ici uniquement l'ordre de grandeur des bornes supérieures obtenues sur l'erreur de reconstruction sans expliciter tous les détails techniques et constantes. L'objet des chapitres 2 et 3 est d'améliorer ces bornes.

Pour le cas de complétion de données binaires ($(Y_i \in \{0,1\})$), Davenport et al. (2014) supposent que la loi des observations est donnée par une fonction de lien $f : \mathbb{R} \mapsto [0,1]$ par $\mathbb{P}(Y_i = 1|k_i, l_i) = f(\bar{X}_{k_i,l_i})$. Ils considèrent alors l'estimateur solution du problème contraint suivant

$$\hat{X} := \underset{\substack{X \in \mathbb{R}^{m_1 \times m_2} \\ \|X\|_\infty \leq \gamma \\ \|X\|_{\sigma,1} \leq \gamma\sqrt{rm_1m_2}}}{\arg\min} \sum_{i=1}^{n} f(X_{k_i,l_i})^{Y_i}(1 - f(X_{k_i,l_i}))^{1-Y_i} ,$$

avec r et γ choisis tels que la vraie matrice $\bar{X}$ appartienne à l'ensemble de contraintes: $\|\bar{X}\|_\infty \leq \gamma$ et $\|X\|_{\sigma,1} \leq \gamma\sqrt{rm_1m_2}$. Sous certaines hypothèses sur l'échantillonnage et la fonction de lien, ils montrent qu'avec grande probabilité, l'estimateur solution de (1.1.3) vérifie

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1m_2} \leq \mathcal{O}\left(\sqrt{\frac{\mathrm{rk}(\bar{X})(m_1 + m_2)}{n}}\right) . \tag{1.10}$$

Gunasekar et al. (2014) se sont intéressés aux cas de distributions appartenant à la famille exponentielle naturelle. Ainsi, il supposent que la densité des données admet la forme suivante:

$$\mathbb{P}(Y|k_i, l_i) \sim h(Y) \exp\left(Y\bar{X}_{k_i,l_i} - G(\bar{X}_{k_i,l_i})\right)$$

avec h la mesure de base et G la fonction de log-partition de la distribution considérée. Ce modèle est suffisamment riche pour englober les modèle gaussiens logistiques entre autres. Pour estimer $\bar{X}$ ils considèrent la log-vraisemblance et résolvent le problème suivant:

$$\hat{X} := \underset{\substack{X \in \mathbb{R}^{m_1 \times m_2} \\ \|X\|_\infty \leq \gamma}}{\arg\min} \sum_{i=1}^{n} \left(G(\bar{X}_{k_i,l_i}) - Y\bar{X}_{k_i,l_i}\right) + \lambda\|X\|_{\sigma,1} ,$$

avec γ choisi de sorte que $\|\bar{X}\|_{\sigma,\infty} \leq \gamma$. Lorsque l'échantillonnage est i.i.d. et uniforme, en choisissant correctement le paramètre de pénalité λ et pour $m_1 \leq m_2$, ils montrent qu'avec grande probabilité on a

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1m_2} \leq \mathcal{O}\left(\max(m_1m_2\gamma^2, 1)\frac{\mathrm{rk}(\bar{X})m_2\log(m_2)}{n}\right) . \tag{1.11}$$

Discussion et problématiques

Comme nous venons de le voir, de nombreux résultats théoriques donnent une bonne supérieure à l'erreur de reconstruction de différentes méthodes de complétion de matrice. On peut légitimement se demander s'il est possible d'améliorer ces garanties et donc de chercher des bornes inférieures.

Intuitivement, on peut répondre à cette question dans le cas bruité en remarquant que si la matrice $\bar{X}$ est de rang r, alors le nombre de degrés de libertés pour le problème de reconstruction est $\mathcal{O}(r(m_1 + m_2))$ et que par conséquent on ne peut améliorer l'erreur au delà de $\mathcal{O}(r(m_1 + m_2)/n)$. De manière plus rigoureuse, définissons l'ensemble $\mathcal{A}(r, \gamma)$ des matrice de rang au plus r et dont l'amplitude des entrées est bornée par γ:

$$\mathcal{F}(r, \gamma) := \{X \in \mathbb{R}^{m_1 \times m_2} |\ \mathrm{rk}(X) \leq r,\ \|X\|_\infty \leq \gamma\} . \tag{1.12}$$

On peut alors énoncer le résultat suivant (voir (Koltchinskii et al., 2011, Théorème 5)) qui confirme l'intuition précédente.

Théorème 3. *Fixons $\gamma > 0$ et $r \in \mathbb{N}$ tel que $1 \leq r \leq \min(m_1, m_2)$ et $n \geq r \max(m_1, m_2)$. Considérons le modèle avec bruit additif (1.1) et supposons que les bruits $\epsilon_1, \cdots, \epsilon_n$ soient i.i.d. gaussiens $\mathcal{N}(0, \sigma^2)$ et que l'échantillonnage soit i.i.d. uniforme et indépendant du bruit. Alors il existe des constantes numériques $c > 0$ et $\beta \in (0, 1)$, telles que pour tout estimateur $\hat{X}$ (fonction des observation à valeur dans $\mathbb{R}^{m_1 \times m_2}$), on ait:*

$$\sup_{X \in \mathcal{F}(r,\gamma)} \mathbb{P}\left(\frac{\|\hat{X} - X\|_{\sigma,2}^2}{m_1 m_2} > c \min(\sigma, \gamma)^2 \frac{\max(m_1, m_2) r}{n} \right) \geq \beta \, .$$

Des résultats similaires peuvent être obtenus pour les problèmes de reconstruction exacte, et l'on peut montrer qu'il faut un échantillon de taille au moins $\propto \mu_0(m_1 + m_2)\log(m_1 + m_2)$ pour espérer reconstruire les matrices de cohérence au plus μ_0 avec grande probabilité (Candès & Tao, 2010, Théorème 1.7).

Au vu de ces résultats, on peut donc constater que les bornes obtenues pour la reconstruction exacte (Théorème 1) et inexacte avec bruit additif sous-exponentiel (Théorème 2) sont optimaux à un facteur logarithmique près. *A contrario*, les résultats concernant les modèles généraux sont loin de la borne intuitive. En effet, pour la complétion de données binaires (1.10), la dépendance par rapport à la taille de l'échantillon est en $1/\sqrt{n}$ contre $1/n$. En ce qui concerne les distributions appartenant à la famille exponentielle (1.11), la dépendance dimensionnelle est en $m_1 m_2 \max(m_1, m_2)$ contre $m_1 + m_2$.

Dans cette thèse, nous nous sommes donc demandé si cette apparente sous optimalité était due à la nature même de ces problèmes ou si les bornes supérieures existantes (1.10) et (1.11) pouvaient être abaissées. Nous avons donc cherché à améliorer les bornes inférieures et supérieures pour les problèmes de complétion binaire (Chapitre 2) et avec distributions appartenant à la famille exponentielle (Chapitre 3). En outre, nous avons cherché à étendre le modèle binaire au cas multinomial (Chapitre 2).

1.1.4 Méthodes d'optimisation pour la complétion de matrice

Les différents estimateurs présentés en Section 1.1.2 ont tous le point commun d'être solution d'un problème de minimisation. Dans cette section, nous présentons les différents algorithmes d'optimisation couramment utilisés en pratique permettant de les calculer. Enfin nous introduisons les problématiques associées que nous avons choisies d'aborder dans cette thèse.

Projections alternées

Pour résoudre le problème (1.3), l'Algorithme Algorithm1 de projections alternées est extrêmement utilisé en pratique (voir Bell & Koren (2007)) à cause de sa simplicité.

Algorithm 1 Projections alternées

1: **Initialisation:** U_0, V_0
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad U_t = \arg\min_{U \in \mathbb{R}^{m_1 \times R}} \sum_{i=1}^n (Y_i - (UV_{t-1}^\top)_{k_i,l_i})^2$
4: $\quad V_t = \arg\min_{V \in \mathbb{R}^{m_1 \times R}} \sum_{i=1}^n (Y_i - (U_t V^\top)_{k_i,l_i})^2$
5: **end for**
6: **Retour:** $X = U_T V_T^\top$.

Cet algorithme nécessite de résoudre des systèmes à $m_1 R$ et $m_2 R$ inconnues à chaque itération ce qui peut se faire de manière très efficace. En outre, il n'est jamais nécessaire de stocker la matrice $\hat{X}$ en mémoire mais seulement ses facteurs U et V. Pourvu que la matrice $\bar{X}$ soit véritablement de rang R, et sous certaines conditions, on peut obtenir des vitesses de convergence vers $\bar{X}$ pour l'Algorithme 1 (voir Keshavan (2012); Jain et al. (2013); Hardt (2013)). Néanmoins, une des faiblesses de cet algorithme est qu'en général le rang de la matrice $\bar{X}$ est inconnu.

Descente de gradient stochastique

Les problèmes de la forme (1.6) permettent de garder la simplicité de la formulation factorisée tout en conservant un aspect adaptatif grâce à la pénalité. La fonction objectif du problème (1.6) que l'on notera F est différentiable et peut donc être minimisée par descente de gradient (*e.g.,* Nesterov (2004)). Lorsque n est grand, le calcul exact du gradient de l'objectif peut s'avérer coûteux et il peut être avantageux d'approximer le gradient à la place (Bottou (2010)). Cela donne lieu à l'algorithme du gradient stochastique (Robbins & Monro (1951)) dont le schéma d'itération s'écrit:

$$\theta_{t+1} = \theta_t - \gamma_t \hat{\nabla} F(\theta_t) \ ,$$

avec $\theta_t := (U_t, V_t)$, $\hat{\nabla} F(\theta_t)$ une approximation du gradient de l'objectif $\nabla F(\theta_t)$ et γ_t le pas d'itération. Pour les problèmes de machine learning dont l'objectif s'écrit comme une somme de nombreux termes comme le problème (1.6), il est naturel de choisir comme approximation le gradient calculé sur un sous échantillon de termes choisis aléatoirement. L'Algorithme 2 décrit l'algorithme de gradient stochastique appliqué au problème (1.6) avec un sous échantillon de taille 1.

Algorithm 2 Gradient stochastique

1: **Initialisation:** Paramètre initial: $X_0 = U_0 V_0^\top$, Suite de pas: $\gamma_1, \gamma_2, \cdots$
2: **for** $t = 1, \ldots, T$ **do**
3: Choisir $i_t \in [n]$:
4: Pour $k \in [m_1]$

$$(U_t)_{k\cdot} = (U_{t-1})_{k\cdot} - \gamma_t \left(\sum_{i=1}^{n} \delta_{k_i,k} (Y_i - (U V_{t-1}^\top)_{k_i,l_i}) (V_{t-1})_{l_i\cdot} + \lambda (U_{t-1})_{k\cdot} \right)$$

5: Pour $l \in [m_2]$:

$$(V_t)_{l\cdot} = (V_{t-1})_{l\cdot} - \gamma_t \left(\sum_{i=1}^{n} \delta_{l_i,l} (Y_i - (U V_{t-1}^\top)_{k_i,l_i}) (U_{t-1})_{k_i\cdot} + \lambda (V_{t-1})_{l\cdot} \right)$$

6: **end for**
7: **Retour:** $X = U_T V_T^\top$.

Il existe une vaste littérature étudiant la vitesse de convergence du gradient stochastique lorsque la fonction objectif à minimiser est convexe (*e.g.,* Nesterov (2004)). Cependant, le problème (1.6) n'est pas convexe et ces résultats ne peuvent s'appliquer directement. Néanmoins, Burer & Monteiro (2005) montrent que si le rang de la solution est inférieur à R, alors les minima locaux sont aussi des minima globaux. Enfin, notons que Recht & Ré (2013) ont adapté l'Algorithme 2 pour proposer une version parallélisable (*Jellyfish*) qui est actuellement l'état de l'art pour les grandes bases de données.

Gradient proximal

Les problèmes (1.4) et (1.8) font apparaître la norme trace qui n'est pas différentiable. Ils peuvent être minimisés par un algorithme proximal (voir Combettes & Pesquet (2011); Parikh et al. (2013)). Pour un problème de la forme

$$\min_{\theta} F(\theta) = f(\theta) + g(\theta) \tag{1.13}$$

avec f une fonction différentiable, l'itération de l'algorithme de gradient proximal s'écrit

$$\theta_{t+1} = \mathrm{prox}_{\gamma_t g}\left(\theta_t - \gamma_t \nabla f(\theta_t)\right) \ ,$$

où prox est l'opérateur proximal. Rappelons que pour une fonction h, l'opérateur proximal prox_h est défini par:

$$\mathrm{prox}_h : \theta \mapsto \arg\min_{\theta'} \frac{1}{2}\|\theta - \theta'\|_2^2 + h(\theta') \ .$$

Par ailleurs, il existe une version accélérée de l'algorithme de gradient proximal (FISTA) proposée par Beck & Teboulle (2009a) qui nécessite de choisir correctement la suite de pas $(\gamma_t)_{t\geq 0}$ et de travailler sur une combinaison de θ_{t+1} et θ_t au lieu de θ_t.

Afin d'appliquer les algorithmes proximaux aux problèmes de complétion, il reste à expliciter l'opérateur proximal associé à la norme trace. Pour un scalaire $\alpha > 0$, appliquer l'opérateur proximal $\mathrm{prox}_{\alpha\|\cdot\|_{\sigma,1}}(X)$ à une matrice X revient à faire un seuillage doux de seuil α des valeurs singulières de X sans en changer les vecteurs singuliers (voir Cai et al. (2010)). L'algorithme de gradient proximal appliqué au problème (1.8) est détaillé ci dessous.

Algorithm 3 Gradient proximal

1: **Initialisation:** X_0, Suite de pas: $\gamma_1, \gamma_2, \cdots$
2: **for** $t = 1, \ldots, T$ **do**
3: SVD: $U\,\mathrm{diag}(\sigma_i)V^\top = X_{t-1} - \gamma_t \nabla L(X_{t-1})$
4: $X_t = U\,\mathrm{diag}(\max(\sigma_i - \gamma_t\lambda, 0))V^\top$
5: **end for**
6: **Retour:** X_T.

Lorsque la fonction d'attache aux données est convexe, alors l'Algorithme 3 converge vers un minimum global.

De nombreux résultats existent pour les versions stochastiques des algorithmes proximaux (voir Hu et al. (2009); Beck & Teboulle (2009b); Juditsky & Nemirovski (2012a,b); Parikh et al. (2013); Xiao & Zhang (2014); Rosasco et al. (2014); Atchade et al. (2014)), et typiquement, ces méthodes convergent en $\mathcal{O}(1/t)$ pour des objectifs fortement convexes (voir Rosasco et al. (2014)). Cependant, ceux-ci sont rarement utilisés pour les problèmes de complétion où n est grand. En effet, pour les problèmes de complétion, on a en général $n \ll m_1 m_2$ et par conséquent le coût de décomposition en valeures singulières (effectuée à chaque itération) est plus important que le coût du calcul exact du gradient. En fait, dans le contexte de la complétion de matrice, le passage à l'échelle des algorithmes proximaux lorsque m_1 et m_2 augmentent, constitue leur principale limitation. En effet, d'une part la SVD devient vite impraticable à cause de sa complexité prohibitive, d'autre part, le coût mémoire pour stocker l'itéré est important.

Algorithme de Frank Wolfe

L'algorithme de Frank Wolfe (aussi connu sous le nom de gradient conditionnel) introduit par Frank & Wolfe (1956) permet de minimiser une fonction convexe f sur un ensemble convexe $\mathcal{C}$:

$$\min_{\theta \in \mathcal{C}} f(\theta) \ . \tag{1.14}$$

L'intérêt de cet algorithme est qu'il ne nécessite pas de calculer une projection à chaque itération contrairement à l'algorithme de gradient projeté.

Algorithm 4 Gradient conditionnel

1: **Initialisation:** θ_0
2: **for** $t = 1, \dots, T$ **do**
3: Résoudre le problème linéaire:

$$a_t = \arg\min_{a \in \mathcal{C}} \langle \nabla f(\theta_t), a \rangle \tag{1.15}$$

4: $\theta_t = \gamma_t a_t + (1 - \gamma_t)\theta_{t-1}$ avec $\gamma_t := 2/(1+t)$
5: **end for**
6: **Retour:** θ_T.

A chaque itération on trouve un minimiseur a_t du problème linéaire contraint (1.15). Cette solution est nécessairement un point extrémal de $\mathcal{C}$ puisque cet ensemble est convexe et l'on appelle a_t un atome. On met ensuite à jour l'itéré en prenant une combinaison convexe de cet atome et de l'itéré précédent. Cette itération est illustrée en Figure 1.2

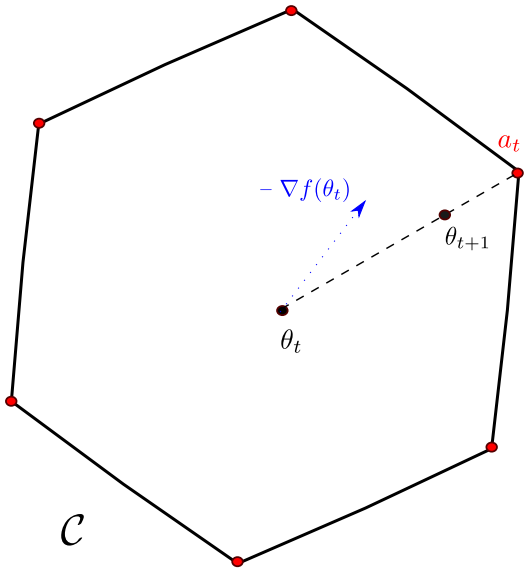

FIGURE 1.2: Mise à jour de l'itéré dans l'algorithme de Frank Wolfe

L'algorithme de Frank Wolfe a récemment regagné en popularité, car le problème (1.15) admet une solution simple à calculer pour beaucoup de problèmes intéressants de machine learning (voir Jaggi (2013)), ce qui en fait une alternative efficace au gradient projeté. En particulier, pour appliquer cet algorithme à la version contrainte du problème (1.8):

$$\min_{X \in \mathbb{R}^{m_1 \times m_2}} L(X) \quad \text{s.c.} \quad \|X\|_{\sigma,1} \leq \rho \ , \tag{1.16}$$

il faut savoir résoudre les problèmes linéaires du type

$$\underset{A\in\mathbb{R}^{m_1\times m_2}}{\arg\min}\ \langle B, A\rangle \quad \text{s.c.} \quad \|A\|_{\sigma,1} \leq \rho\ , \tag{1.17}$$

pour n'importe quelle matrice B. Il est aisé de voir, qu'une solution de (1.17) est donnée par $A = \rho uv^\top$ avec u (*resp.* v) le vecteur singulier de gauche (*resp.* de droite) associé à la plus grande valeur singulière de B. L'Algorithme 5 détaille l'application de l'algorithme de Frank Wolfe au problème (1.16).

Algorithm 5 Gradient conditionnel pour le problème (1.16)

1: **Initialisation:** X_0
2: **for** $t = 1, \ldots, T$ **do**
3: Calcul de u_t, v_t les vecteurs singuliers associés à la plus grande valeur singulière de $\nabla f(X_t)$
4: $X_t = \gamma_t u_t v_t^\top + (1-\gamma_t)X_{t-1}$ avec $\gamma_t := 2/(1+t)$
5: **end for**
6: **Retour:** θ_T.

L'algorithme de Frank Wolfe présente deux avantages par rapport au gradient proximal pour les problèmes de complétion. Premièrement, il ne nécessite de calculer que les vecteurs singuliers associés à la première valeur singulière du gradient. La parcimonie de ce dernier ($n \ll m_1 m_2$) combiné à des algorithmes de puissances itérées tel que l'algorithme de Lanczos (voir par exemple Golub & van Loan (1996)), permet de rendre ce calcul très efficace. Enfin, il n'est pas nécessaire de garder en mémoire la totalité de la matrice X_t pour faire les calculs. En effet, la liste des valeurs des entrées de X_t pour les coefficients échantillonnés suffit. Hazan & Kale (2012) ont proposé une version stochastique de l'algorithme de Frank Wolfe mais les garanties de convergence obtenues sont sensiblement moins bonnes que pour les algorithmes de gradient stochastique ($\mathcal{O}(\sqrt{\log(t)/t})$ contre $\mathcal{O}(1/t)$ pour un objectif fortement convexe et lisse). Nous revenons en détail sur les versions stochastiques de l'algorithme de Frank Wolfe au Chapitre 4.

Discussions et problématiques

Comme nous venons de le voir, on peut distinguer trois familles de méthodes d'optimisation pour la complétion de matrice.

Il y a d'abord les algorithmes qui s'intéressent aux formulations factorisées (1.3) et (1.6). En travaillant sur les facteurs U et V plutôt que sur la matrice X, les problèmes deviennent différentiables et peuvent être minimisés par des algorithmes efficaces et simples (projections alternées et gradient stochastique) qui passent à l'échelle. L'inconvénient est que pour l'instant, il n'existe de preuve de convergence vers un minimum global que lorsque la fonction de perte est quadratique. Ainsi, il n'est pas possible de les généraliser aux problèmes (1.8) sans perdre leurs garanties théoriques.

Deuxièmement, il y les algorithmes de type gradient proximal (ou gradient projeté pour les problèmes contraints). Il existe des garanties sur les vitesses de convergence vers un minimum global pour ces algorithmes dès que la fonction d'attache aux données est convexe. Ces garanties existent aussi pour les versions stochastiques. Néanmoins comme ils requièrent de faire une décomposition en valeur singulière à chaque itération, ces algorithmes sont peu adaptés aux problèmes de grande dimension *i.e.,* pour $m_1, m_2 \gg 1$.

Enfin, les algorithmes de gradient conditionnel garantissent la convergence vers un minimum global pour les objectifs convexes et sont adaptés aux problèmes de grande dimension. Cependant, les garanties de convergence pour les versions stochastiques restent sensiblement moins bonnes que celles des algorithmes de gradients stochastique.

Nous avons choisi de nous intéresser à ce dernier point dans cette thèse et cherché sous quelles conditions, les algorithmes de gradient conditionnels stochastiques sont capables de garantir les mêmes vitesses de convergences que les algorithmes de gradient stochasitque.

1.2 Contributions

1.2.1 Complétion de matrice pour les modèles multinomiaux

1.2.1.1 Résumé

Dans le Chapitre 2, on s'intéresse au problème de complétion de matrice lorsque les observations ne peuvent prendre qu'un nombre fini de valeurs. Plus précisément, on considère le schéma d'observations $(Y_i, k_i; l_i)_{i=1}^n$ avec $k_i \in [m_1]$, $l_i \in [m_2]$ et $Y_i \in \{1, \cdots, p\}$. Le but est de pouvoir traiter les problèmes où les données sont naturellement discrètes ou catégoriques (*e.g.,* quantisation, sondage).

Cas binaire On considère d'abord le cas de données binaires ($p = 2$) qui avait été étudié préalablement par Davenport et al. (2014). On suppose que les observations sont indépendantes et suivent une distribution logistique paramétrée par une matrice $\bar{X}$ de la façon suivante:

$$\mathbb{P}(Y_i = j) = f^j(\bar{X}_{k_i,l_i}), \quad j \in \{1, 2\} \, , \tag{1.18}$$

où $f := (f^j)_{j=1}^2$ est une fonction de lien *i.e.,* $f^j \geq 0$ et $f^1 + f^2 = 1$. On cherche à estimer $\bar{X}$ à partir des observations.

Etant donné une borne supérieur $\gamma > 0$ de l'amplitude maximale des entrées du paramètre $\|\bar{X}\|_\infty$, on considère l'estimateur suivant:

$$\hat{X} = \underset{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_\infty \leq \gamma}{\arg\min} \Phi(X) \, , \quad \text{où} \quad \Phi(X) = \Phi_{\mathrm{Y}}(X) + \lambda \|X\|_{\sigma,1} \, , \tag{1.19}$$

avec $\Phi_{\mathrm{Y}}(\cdot)$ l'opposé de la log-vraisemblance conditionnelle (à l'échantillonnage) du modèle (1.18) et $\lambda > 0$ un hyper paramètre que l'on fixera ultérieurement.

Le premier résultat obtenu est une borne supérieure sur le risque de reconstruction. En supposant que l'échantillonnage des entrées satisfasse H4, on peut montrer qu'en choisissant correctement λ, on a avec grande probabilité (voir Corollaire 2.3) :

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \max\left(C_1 \frac{\mathrm{rk}(\bar{X}) \max(m_1, m_2) \log(m_1 + m_2)}{n}, C_2 \sqrt{\frac{\log(m_1 + m_2)}{n}} \right) \, .$$

avec C_1 et C_2 des constantes qui dépendent de la distribution d'échantillonnage, de γ et de la fonction de lien f. Ce résultat constitue une amélioration significative par rapport à la borne supérieure (1.10) obtenue par Davenport et al. (2014). Une borne analogue est donnée pour l'erreur de reconstruction mesurée par la divergence de Kullback Leibler plutôt qu'en norme de Frobenius (voir Théorème 2.2).

Le second résultat donne une borne inférieure sur l'erreur de reconstruction. Il montre que sous certaines hypothèses sur la fonction de lien (voir Théorème 2.5) il existe $\beta > 0$ et $c > 0$ tels que pour tout $m_1, m_2 \geq 2$, $1 \leq r \leq \min(m_1, m_2)$, et $\gamma > 0$ on a

$$\inf_{\hat{X}} \sup_{\bar{X} \in \mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}} \left(\frac{\|\hat{X} - \bar{X}\|_2^2}{m_1 m_2} > c \min \left\{ \gamma^2, \frac{\max(m_1, m_2) r}{n} \right\} \right) \geq \beta \, ,$$

où $\mathcal{F}(r, \gamma)$ a été défini en (1.12). Ainsi, sous certaines conditions, la borne supérieure obtenue précédemment est optimale par rapport à m_1, m_2 et n à un facteur logarithmique près. Davenport et al. (2014) ont aussi dérivé une borne inférieure, néanmoins ils ne considère pas l'ensemble $\mathcal{F}(r, \gamma)$ mais $\mathcal{A}(r, \gamma)$ défini par

$$\mathcal{A}(r, \gamma) := \{X \in \mathbb{R}^{m_1 \times m_2} | \ \|X\|_{\sigma,1} \leq \gamma \sqrt{r m_1 m_2}, \ \|X\|_\infty \leq \gamma\} \, .$$

Remarquons que l'on a $\mathcal{F}(r, \gamma) \subset \mathcal{A}(r, \gamma)$ et que ceci explique pourquoi la borne inférieure obtenue par Davenport et al. (2014) est différente de celle du Théorème 2.5.

Cas général Le cas binaire est ensuite étendu au cas $p \geq 2$. Dans ce cas-là, la distribution des observations n'est pas paramétrée par une matrice mais par un tenseur $\bar{\mathcal{X}} := (\bar{X}^l)_{l=1}^q$ de la façon suivante:

$$\mathbb{P}(Y_i = j) = f^j(\bar{\mathcal{X}}_{k_i,l_i}), \quad j \in [p] \, , \tag{1.20}$$

où $f = (f^j)_{j=1}^p$ est une fonction de lien et $\bar{\mathcal{X}}_{k_i;l_i}$ est le vecteur $(\bar{X}^l_{k_i,l_i})_{l=1}^q$. Il est important de noter qu'il y a ici deux dimensions en jeu: p et q. Ainsi, chaque valeur Y_i appartient à l'ensemble $\{1, \cdots, p\}$ et admet une distribution paramétrée par un vecteur de dimension q: $(\bar{X}^l_{k_i,l_i})_{l=1}^q$. Par exemple, dans le cas d'une distribution logistique, on aura $q = p - 1$ et

$$f^j(\bar{\mathcal{X}}_{\omega_i}) = \frac{\exp(\bar{X}^j_{k_i,l_i})}{1 + \sum_{l=1}^{p-1} \exp(\bar{X}^l_{k_i,l_i})} \quad \text{pour} \quad j \in [p-1] \, .$$

Afin de reconstruire le tenseur $\bar{\mathcal{X}}$ on considère l'estimateur suivant:

$$\hat{\mathcal{X}} = \underset{\substack{\mathcal{X} \in \mathbb{R}^{m_1 \times m_2 \times q} \\ \|\mathcal{X}\|_\infty \leq \gamma}}{\arg\min} \Phi(\mathcal{X}) \, , \quad \text{where} \quad \Phi(\mathcal{X}) = \Phi_{\mathrm{Y}}(\mathcal{X}) + \lambda \sum_{j=1}^q \|X^j\|_{\sigma,1} \, , \tag{1.21}$$

avec $\Phi_{\mathrm{Y}}(\cdot)$ l'opposé de la log-vraisemblance du modèle (1.20).

En supposant que la fonction de lien se factorise comme

$$f^j(x_1, \ldots, x_q) = \prod_{l=1}^q g_l^j(x_l) \quad \text{for } j \in [p] \, ,$$

et sous H4, on peut majorer le risque de reconstruction avec grande probabilité (voir Théorème 2.7) et obtenir une borne supérieure en

$$\mathcal{O}\left(\max\left(q^2 \frac{\mathrm{rk}(\bar{X}) \max(m_1, m_2) \log(m_1 + m_2)}{n}, \sqrt{\frac{\log(m_1 + m_2)}{n}} \right) \right) \, .$$

La dépendance par rapport à m_1, m_2 et n est la même que pour le cas binaire mais la borne supérieure est proportionnelle à un facteur q^2.

Implémentation Le calcul des estimateurs a été fait en implémentant un algorithme proposé par Dudík et al. (2012); Harchaoui et al. (2012) qui a été étendu pour prendre en compte l'estimateur de tenseur (1.21). Cet algorithme est une adaptation du gradient conditionnel aux versions lagrangiennes des problèmes contraints et il ne nécessite pas de faire une décomposition en valeurs singulières à chaque itération (contrairement aux algorithmes proximaux).

En effet, l'itération de base de cet algorithme consiste à calculer les vecteurs singuliers du gradient de $\Phi_{\mathrm{Y}}(\cdot)$ associés à la première valeur singulière. Ce calcul peut être rendu efficace en prenant avantage de la sparsité de ce gradient (qui est nul partout sauf pour les entrées observées) avec l'algorithme des itération d'Arnoldi (*e.g.,* Golub & van Loan (1996)). Le tableau 1.3 donne l'ordre de grandeur du temps d'exécution pour une implémentation en langage C exécuté avec une machine avec processeur Xeon 3.07Ghz w3550, RAM 1.66 Go, Cache 8 Mo.

Parameter Size	1000×1000	3000×3000	10000×10000
Observations	$100 \cdot 10^3$	$1 \cdot 10^6$	$10 \cdot 10^6$
Execution Time (s.)	4.5	52	730

TABLE 1.3: Execution time of the proposed algorithm for the binary case.

Résumé des contributions et publications associées

Dans cette partie, nous avons donc proposé un estimateur à pénalité norme trace pour les problèmes de complétion avec des données discrètes ou catégoriques. Dans le cas binaire, nous montrons qu'avec grande probabilité cet estimateur est optimal à un facteur logarithmique près en donnant une borne inférieure et supérieure sur l'erreur de reconstruction. Ces résultats améliorent l'état de l'art en réduisant d'un facteur $1/\sqrt{n}$ la borne supérieure donnée par Davenport et al. (2014). Pour le cas général, nous donnons une borne supérieure sur l'erreur de reconstruction. Ceci constitue un premier résultat pour des données multinomiales. Ces travaux ont donné lieu à deux publications: Lafond et al. (2014); Klopp et al. (2014).

1.2.2 Complétion de matrice avec bruit appartenant à la famille exponentielle

1.2.2.1 Résumé

Dans le chapitre 3, nous nous intéressons au cas des problèmes de complétion de matrice lorsque la distribution des observations appartient à la famille exponentielle naturelle. Nous supposons que nous disposons d'observations de la forme $(Y_i, k_i; l_i)_{i=1}^n$ et que la distribution de Y_i conditionnellement à l'échantillonnage appartient à la famille exponentielle et est paramétrée par une matrice $\bar{X}$ de la façon suivante:

$$Y_i | k_i, l_i \sim \mathrm{Exp}_{h,G,(}\, \bar{X}_{k_i,l_i}) := h(Y_i) \exp\left(\bar{X}_{k_i,l_i} Y_i - G(\bar{X}_{\omega_i})\right) \ , \tag{1.22}$$

avec h la mesure de base et G la fonction de log-partition. La famille exponentielle est suffisamment large pour inclure de nombreuses distributions intéressantes pour les applications. Le tableau 1.4 donne une liste non exhaustive de distributions appartenant à la famille exponentielle. Nous cherchons à reconstruire la matrice $\bar{X}$ et nous considérons deux cas. Nous donnons d'abord un estimateur qui ne nécessite pas de connaître la distribution de l'échantillonnage et donnons une borne supérieure sur l'erreur de reconstruction. Ensuite, nous montrons que lorsque la distribution est connue, il existe un autre estimateur avec de

Distribution	Paramètre x	$G(x)$
Gaussian: $\mathcal{N}(\mu, \sigma^2)$ (σ connu)	μ/σ	$\sigma^2 x^2/2$
Binomial: $\mathcal{B}^N(p)$ (N connu)	$\log(p/(1-p))$	$N\log(1+\mathrm{e}^x)$
Poisson: $\mathcal{P}(\lambda)$	$\log(\lambda)$	e^x
Exponential: $\mathcal{E}(\lambda)$	$-\lambda$	$-\log(-x)$

TABLE 1.4: Quelques distributions de la famille exponentielle

meilleures garanties de reconstruction. Enfin nous donnons une borne inférieure sur l'erreur de reconstruction.

Distribution d'échantillonnage inconnue Conditionnellement à l'échantillonnage, l'opposé de la log-vraisemblance du modèle est donné par

$$\Phi_{\mathrm{Y}}(X) = -\frac{1}{n}\sum_{i=1}^{n}\left(\log(h(Y_i)) + X_i Y_i - G(X_i)\right) \ . \tag{1.23}$$

Etant donné une borne supérieure $\gamma > 0$ de l'amplitude maximale des entrées de la matrice $\bar{X}$, on considère dans un premier temps l'estimateur à pénalité norme trace suivant:

$$\hat{X} = \underset{X\in\mathbb{R}^{m_1\times m_2}, \|X\|_\infty \leq \gamma}{\arg\min} \Phi(X) \ , \quad \text{où} \quad \Phi(X) = \Phi_{\mathrm{Y}}(X) + \lambda\|X\|_{\sigma,1} \ , \tag{1.24}$$

avec $\lambda > 0$ un hyper paramètre que l'on fixera ultérieurement.

Le premier résultat obtenu est une borne supérieure sur le risque de reconstruction de cet estimateur. En supposant que l'on échantillonne des entrées vérifie H4, et que la distribution des valeurs Y_i conditionnellement à k_i, l_i soit sous-exponentielle (voir H5) nous montrons qu'en choisissant correctement λ, on a avec grande probabilité (voir Théorème 3.6) :

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \max\left(C_1 \frac{\mathrm{rk}(\bar{X})\max(m_1,m_2)\log(m_1+m_2)}{n}, C_2\sqrt{\frac{\log(m_1+m_2)}{n}}\right) \ .$$

avec C_1 et C_2 des constantes qui dépendent de la distribution d'échantillonnage, de γ et de la fonction de log-partition G. Ce résultat améliore donc la borne (1.11) obtenue Gunasekar et al. (2014) pour un modèle similaire d'un facteur $m_1 m_2$.

Distribution d'échantillonnage connue Lorsque la distribution d'échantillonnage est connue (*e.g.,* lorsque l'on choisi le schéma d'échantillonnage), il est possible de définir l'estimateur suivant

$$\check{X} := \underset{X\in\mathbb{R}^{m_1\times m_2}, \|X\|_\infty \leq \gamma}{\arg\min} \Phi_{\mathrm{Y}}^{\Pi}(X) + \lambda\|X\|_{\sigma,1} \quad \text{with} \ , \tag{1.25}$$

$$\Phi_{\mathrm{Y}}^{\Pi}(X) := G^{\Pi}(X) - \frac{\sum_{i=1}^n X_{k_i,l_i} Y_i}{n} \quad \text{and} \quad G^{\Pi}(X) := \mathbb{E}\left[\frac{\sum_{i=1}^n G(X_{k_i,l_i})}{n}\right] \ ,$$

en effet, le terme d'espérance $G^{\Pi}(\cdot)$ nécessite de connaître la distribution de $(k_i, l_i)_{i=1}^n$. Nous montrons que cet estimateur satisfait une inégalité oracle. Avant d'énoncer le résultat, pour toute matrice $X^1, X^2 \in \mathbb{R}^{m_1\times m_2}$ notons $D_G(X^1, X^2)$ la divergence de Kullback entre les distributions des observations $(Y_i, k_i, l_i)_{i=1}^n$ paramétrée par X^1 et X^2 (voir le chapitre 3 pour une définition complète).

Alors, sous l'hypothèse $H4$ et pour $\lambda \geq \|\nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})\|_{\sigma,\infty}$, nous montrons qu'avec grande probabilité, nous avons les deux inégalités oracles suivantes (voir Théorème 3.9):

$$\frac{D_G(\check{X}, \bar{X})}{m_1 m_2} \leq \inf_{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_\infty \leq \gamma} \left(\frac{D_G(X, \bar{X})}{m_1 m_2} + 2\frac{\lambda \|X\|_{\sigma,1}}{m_1 m_2} \right) \tag{1.26}$$

et

$$\frac{D_G(\check{X}, \bar{X})}{m_1 m_2} \leq \inf_{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_\infty \leq \gamma} \left(\frac{D_G(X, \bar{X})}{m_1 m_2} + C\lambda^2 \operatorname{rk}(X) \right) , \tag{1.27}$$

avec C une constante dépendant de l'échantillonnage et de G. Nous donnons ensuite des conditions afin de contrôler λ (qui dépend de la quantité aléatoire $\|\nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})\|_{\sigma,\infty}$) avec grande probabilité. Ce résultat est important car il permet de mesurer l'erreur de reconstruction même lorsque le vrai paramètre $\bar{X}$ ne satisfait pas , $\|\bar{X}\|_\infty \leq \gamma$. Ce résultat étend les travaux de Koltchinskii et al. (2011).

Borne inférieure Enfin, nous donnons une borne inférieure sur l'erreur de reconstruction pour le modèle (1.22). Nous montrons qu'il existe des constantes $c > 0$ dépendant de la fonction de log-partition G et $\beta > 0$ telles que pour tout $m_1, m_2 \geq 2$, $1 \leq r \leq \min(m_1, m_2)$ et $\gamma > 0$ on a

$$\inf_{\hat{X}} \sup_{\bar{X} \in \mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}} \left(\frac{\|\hat{X} - \bar{X}\|_2^2}{m_1 m_2} > c \min \left\{ \gamma^2, \frac{\max(m_1, m_2) r}{n} \right\} \right) \geq \beta ,$$

où $\mathcal{F}(r, \gamma)$ a été défini en (1.12). En d'autres termes, les bornes supérieures obtenues précédemment sont optimales par rapport à m_1, m_2 et n à un facteur logarithmique près.

Résumé des contributions et publications associées

Dans le chapitre 3 nous considérons le problème de complétion de matrice lorsque les observations suivent une distribution appartenant à la famille exponentielle. Nous considérons un premier estimateur à pénalité norme trace et majorons avec grande probabilité l'erreur de reconstruction mesurée en norme de Frobenius. La borne obtenue améliore d'un facteur $m_1 m_2$ celle précédemment obtenue par Gunasekar et al. (2014). Lorsque le schéma d'échantillonnage des entrées est connu nous considérons un second estimateur et prouvons une inégalité oracle sur l'erreur de reconstruction mesurée en divergence de Kullback Leibler. Enfin, nous donnons une borne inférieure sur l'erreur de reconstruction et montrons que les garanties obtenues sont optimales à un facteur logarithmique près. Ces résultats ont été publiés (voir Lafond (2015)).

1.2.3 Algorithmes de Frank Wolfe stochastiques

1.2.3.1 Résumé

Dans le chapitre 4, nous nous intéressons à des versions stochastiques de l'algorithme du gradient conditionnel. Comme mentionné en introduction, cet algorithme est intéressant pour beaucoup d'applications (dont la complétion de matrice) en machine learning car il permet de minimiser des problèmes contraints sans nécessiter de faire de projection. Les versions stochastiques du gradient conditionnel ont été beaucoup moins étudiées que les algorithmes de descente de gradient. Hazan & Kale (2012) ont considéré une version stochastique de l'algorithme de Frank Wolfe et montré que la vitesse de convergence vers le minimum est

majorée par $\mathcal{O}(\sqrt{\log^2 t/t})$ pour les objectifs fortement convexes (contre $\mathcal{O}(1/t)$ pour le gradient stochastique). Cette garantie a été améliorée par Garber & Hazan (2015a,b); Lan & Zhou (2014); Hazan & Luo (2016) qui ont considéré des variantes d'algorithme de descente de gradient projetés où la projection est approximée en utilisant Frank Wolfe. Ils atteignent une vitesse de convergence en $\mathcal{O}(\mathrm{poly}(\log(t))/t)$ avec (poly un polynôme de degré au plus 3. Néanmoins, pour implémenter ces algorithmes il faut connaître des informations sur l'objectif telles qu'une borne supérieure sur le gradient ou une borne inférieure sur la forte convexité. Ceci est un désavantage pour bon nombre de problèmes d'apprentissage stochastique où ces bornes sont inconnues. Dans ce travail, nous présentons des versions stochastiques simple (*i.e.,* ne nécessitant pas d'information *a priori* sur l'objectif pour être implémentées) du gradient conditionnel et donnons des hypothèses sous lesquelles des garanties de convergence en $\mathcal{O}(\mathrm{poly}(\log(t))/t)$ peuvent être apportées. Nous étudions la convergence de ces algorithmes pour des objectifs non convexes et proposons des applications pour l'apprentissage en ligne.

Algorithmes Nous considérons le problème d'optimisation suivant

$$\arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \;\text{ s.t. }\; \boldsymbol{\theta} \in \mathcal{C} \;, \tag{1.28}$$

avec $\mathcal{C}$ un ensemble de contraintes convexe et borné. Nous nous intéressons au cas où à chaque itération t, on ne peut pas calculer exactement le gradient de l'objectif $f(\cdot)$ mais seulement une approximation $\hat{\nabla}_t f(\cdot)$. Nous étudions l'algorithme de Frank Wolfe où le gradient exact est remplacé par cette approximation (S-FW voir l'Algorithme 6). La séquence de pas γ_t à choisir est donné ultérieurement.

Algorithm 6 Frank-Wolfe Stochastique (S-FW).

1: **Initialize:** $\boldsymbol{\theta}_1 \leftarrow 0$
2: **for** $t = 1, \ldots$ **do**
3: Résoudre:

$$\boldsymbol{a}_t \leftarrow \arg\min_{\boldsymbol{a} \in \mathcal{C}} \left\langle \boldsymbol{a}, \hat{\nabla}_t f(\boldsymbol{\theta}_t) \right\rangle . \tag{1.29}$$

4: $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \gamma_t(\boldsymbol{a}_t - \boldsymbol{\theta}_t)$.
5: **end for**

On considère aussi la version stochastique de l'algorithme de Frank Wolfe avec pas de côté ("away step") introduite par Wolfe (1970). Par convexité de l'ensemble de contrainte, l'itéré $\boldsymbol{\theta}_t$ est une combinaison convexe de points extrémaux de $\mathcal{C}$ que nous appelons atomes. Nous définissons $\mathcal{A}_t$ l'ensemble de ces atomes actifs (*i.e.,* intervenant dans la combinaison avec un poids strictement positif) et notons $\alpha_t^{\boldsymbol{a}} > 0$ le poids de chaque atome actif $\boldsymbol{a} \in \mathcal{A}_t$ à l'itération t:

$$\boldsymbol{\theta}_t = \textstyle\sum_{\boldsymbol{a} \in \mathcal{A}_t} \alpha_t^{\boldsymbol{a}} \cdot \boldsymbol{a} \quad \text{avec} \quad \alpha_t^{\boldsymbol{a}} > 0 \;. \tag{1.30}$$

L'algorithme de 'away step' stochastique (S-AW voir l'Algorithme 7) est similaire à l'algorithme S-FW, si ce n'est qu'il est possible lors de certaines itération de réduire la contribution d'un atome actif et d'augmenter celle des autres (voir lignes 9 et 11). On parle alors "d'away step" pour ces itérations.

Algorithm 7 Algorithme de Frank Wolfe avec Away Step (S-AW).

Initialize: $n_0 = 0$, $\boldsymbol{\theta}_1 = 0$, $\mathcal{A}_1 = \emptyset$;
for $t = 1, \dots$ **do**
  Résoudre les problèmes linéaires suivants:

$$\boldsymbol{a}_t^{\mathsf{FW}} \leftarrow \arg\min_{\boldsymbol{a} \in \mathcal{C}} \left\langle \boldsymbol{a}, \hat{\nabla}_t f(\boldsymbol{\theta}_t) \right\rangle, \; \boldsymbol{a}_t^{\mathsf{AW}} \leftarrow \arg\max_{\boldsymbol{a} \in \mathcal{A}_t} \left\langle \boldsymbol{a}, \hat{\nabla}_t f(\boldsymbol{\theta}_t) \right\rangle \tag{1.31}$$

  if $\langle \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t, \hat{\nabla}_t f(\boldsymbol{\theta}_t) \rangle \leq \langle \boldsymbol{\theta}_t - \boldsymbol{a}_t^{\mathsf{AW}}, \hat{\nabla}_t f(\boldsymbol{\theta}_t) \rangle$ or $\mathcal{A}_t = \emptyset$ **then**
    `FW step:` $\boldsymbol{d}_t \leftarrow \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t$, $n_t \leftarrow n_{t-1} + 1$, $\hat{\gamma}_t \leftarrow \gamma_{n_t}$ et $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup \{\boldsymbol{a}_t^{\mathsf{FW}}\}$.
  else
    $\boldsymbol{d}_t \leftarrow \boldsymbol{\theta}_t - \boldsymbol{a}_t^{\mathsf{AW}}$, $\gamma_{\max} = \alpha_t^{\boldsymbol{a}_t^{\mathsf{AW}}} / (1 - \alpha_t^{\boldsymbol{a}_t^{\mathsf{AW}}})$, cf. (1.30) pour la définition de $\alpha_t^{\boldsymbol{a}_t^{\mathsf{AW}}}$.
    if $\gamma_{\max} \geq \gamma_{n_t - 1}$ **then**
      `AW step:` $n_t \leftarrow n_{t-1} + 1$ et $\hat{\gamma}_t \leftarrow \gamma_{n_t}$
    else
      `Drop step:` $\hat{\gamma}_t \leftarrow \gamma_{\max}$, $n_t \leftarrow n_{t-1}$ et $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \setminus \{\boldsymbol{a}_t^{\mathsf{AW}}\}$
    end if
  end if
  Compute $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \hat{\gamma}_t \boldsymbol{d}_t$.
end for

Analyse de la convergence Nous donnons des bornes supérieures sur les vitesses de convergence des algorithmes S-FW et S-AW lorsque l'objectif f est fortement convexe et lisse. Nous fixons $\gamma_t = 2/(t+1)$ dans les algorithmes S-FW et S-AW. Nous supposerons que le bruit d'approximation peut être contrôlé et qu'il existe $\alpha \in (0, 1]$, $\sigma \geq 0$ tel qu'avec probabilité au moins $1 - \epsilon$,

$$\|\hat{\nabla}_t f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\| = \mathcal{O}\left(\frac{\eta_t^\epsilon}{t}\right)^\alpha , \tag{1.32}$$

où $\eta_t^\epsilon \geq 1$ est tel que le membre de droite tende vers 0. Pour certains exemples d'apprentissage en ligne nous montrons (voir Proposition 6 du Chapitre 4) que l'hypothèse précédente est vérifiée pour $\alpha = 0.5$ et $\eta_t^\epsilon = \log(t)$. Lorsque l'optimum $\boldsymbol{\theta}^*$ appartient à l'intérieur de l'ensemble de contrainte $\mathcal{C}$, nous montrons alors que pour l'algorithme S-FW on a

$$f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^*) \leq \mathcal{O}\left(\frac{\eta_t^\epsilon}{t}\right)^{2\alpha} . \tag{1.33}$$

Lorsque l'ensemble de contrainte est un polytope nous montrons que la vitesse de convergence de l'algorithme S-AW vérifie aussi (1.33). Ces résultats sont détaillés dans le Théorème 4 et leur application à l'apprentissage en ligne dans le Théorème 7.

Nous nous intéressons aussi à la convergence de ces algorithmes dans le cas d'objectifs non convexes. Pour un itéré $\boldsymbol{\theta}_t$, nous prenons comme mesure de stationnarité le saut de dualité de Frank Wolfe g_t défini par

$$g_t := \max_{\boldsymbol{\theta} \in \mathcal{C}} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta} \rangle .$$

Cette mesure de la stationnarité qui diffère du choix habituel consistant à prendre $\|\nabla f_t(\boldsymbol{\theta}_t)\|$, apparaît naturellement dans les équations dérivées de l'algorithme de Frank Wolfe. En outre, elle présente l'avantage d'être invariante par reparamétrisation affine. Nous montrons que lorsque l'objectif est lisse, borné et que l'hypothèse (1.32) est satisfaite alors on a (voir Théorème 5)

$$g_t = \mathcal{O}\left(\frac{\eta_t^\epsilon}{t}\right)^\alpha .$$

Résumé des contributions et publications associées

Dans le Chapitre 4 nous donnons des versions stochastiques de l'algorithme de Frank Wolfe qui ne nécessitent pas d'information *a priori* sur l'objectif pour être implémentées. En outre, les garanties de convergence de ces algorithmes lorsque l'objectif est fortement convexe et sous certaines hypothèses, atteignent l'Etat de l'art des algorithmes d'optimisation contrainte ne nécessitant pas de projections. Cela est résumé dans le Tableau 1.5.

	Cadre	borne (*regret*)	borne (*stochastique*)
Garber and Hazan, 2015 Garber & Hazan (2015b)	Obj. Lipschitz cvx. *	$\mathcal{O}(\sqrt{1/t})$	$\mathcal{O}(\sqrt{\log t/t})$
	Obj. fort. cvx. *†	$\mathcal{O}(\log t/t)$	$\mathcal{O}(\log t/t)$
Hazan and Kale, 2012 Hazan & Kale (2012)	Obj. Lipschitz cvx.	$\mathcal{O}(\sqrt{\log^2 t/t})$	$\mathcal{O}(\sqrt{\log^2 t/t})$
	Obj. fort. cvx.	$\mathcal{O}(\sqrt{\log^2 t/t})$	$\mathcal{O}(\sqrt{\log^2 t/t})$
Ce travail	Obj. fort. cvx., opt. intérieur (S-FW)	$\mathcal{O}(\log^3 t/t)$	$\mathcal{O}(\log^2 t/t)$
	Obj. fort. cvx., cont. polytope (S-AW)	$\mathcal{O}(\log^3 t/t)$	$\mathcal{O}(\log^2 t/t)$

Nécessite: *borne sup. sur Lipschitz, †borne inf. sur forte convexité

TABLE 1.5: Vitesse de convergence donnée dans le contexte de l'optimisation stochastique et en ligne (regret)

Ces résultats sont disponibles sur arXiv (voir Lafond et al. (2016)) et seront soumis pour une publication journal ultérieurement. Enfin, dans Lafond et al. (2016) nous avons considéré une application des algorithmes précédents à l'apprentissage distribué.

CHAPTER 2
Adaptive multinomial matrix completion

The task of estimating a matrix given a sample of observed entries is known as the *matrix completion problem*. Most works on matrix completion have focused on recovering an unknown real-valued low-rank matrix from a random sample of its entries. Here, we investigate the case of highly quantized observations when the measurements can take only a small number of values. These quantized outputs are generated according to a probability distribution parametrized by the unknown matrix of interest. This model corresponds, for example, to ratings in recommender systems or labels in multi-class classification. We consider a general, non-uniform, sampling scheme and give theoretical guarantees on the performance of a constrained, nuclear norm penalized maximum likelihood estimator. One important advantage of this estimator is that it does not require knowledge of the rank or an upper bound on the nuclear norm of the unknown matrix and, thus, it is adaptive. We provide lower bounds showing that our estimator is minimax optimal. An efficient algorithm based on lifted coordinate gradient descent is proposed to compute the estimator. A limited Monte-Carlo experiment, using both simulated and real data is provided to support our claims.

2.1 Introduction

The matrix completion problem arises in a wide range of applications such as image processing Hui et al. (2010); Ji et al. (2013); Xu et al. (2014), quantum state tomography Gross (2011), seismic data reconstruction Yang et al. (2013) or recommender systems Koren et al. (2009); Bobadilla et al. (2013). It consists in recovering all the entries of an unknown matrix, based on partial, random and, possibly, noisy observations of its entries. Of course, since only a small proportion of entries is observed, the problem of matrix completion is, in general, ill-posed and requires a penalization favoring low rank solutions. In the classical setting, the entries are assumed to be real valued and observed in presence of additive, homoscedastic Gaussian or sub-Gaussian noise. In this framework, the matrix completion problem can be solved provided that the unknown matrix is low rank, either exactly or approximately; see Candès & Plan (2010); Keshavan et al. (2010); Koltchinskii et al. (2011); Negahban & Wainwright (2012); Cai & Zhou (2013a); Klopp (2014) and the references therein. Most commonly used methods amount to solve a least square program under a rank constraint or its convex relaxation provided by the nuclear (or trace) norm Fazel (2002).

In this paper, we consider a statistical model where instead of observing a real-valued entry of an unknown matrix we are now able to see only highly quantized outputs. These discrete observations are generated according to a probability distribution which is parameterized by the corresponding entry of the unknown low-rank matrix. This model is well suited to the analysis of voting patterns, preference ratings, or recovery of incomplete survey data, where typical survey responses are of the form "true/false", "yes/no" or "agree/disagree/no opinion" for instance.
The problem of matrix completion over a finite alphabet has received much less attention than the traditional unquantized matrix completion. One-bit matrix completion, corresponding to the case of binary, i.e. yes/no, observations, was first introduced by Davenport et al. (2012). In this paper, the first theoretical guarantees on the performance of a nuclear-norm constrained maximum likelihood estimator are given. The sampling model considered in Davenport et al. (2012) assumes that the entries are sampled uniformly at random. Unfortunately, this condition is unrealistic for recommender system applications: in such a context some users are more active than others and popular items are rated more frequently. Another important issue is that the method of Davenport et al. (2012) requires the knowledge of an upper bound on the nuclear norm or on the rank of the unknown matrix. Such information is usually not available

in applications. On the other hand, our estimator yields a faster rate of convergence than those obtained in Davenport et al. (2012).

One-bit matrix completion was further considered by Cai & Zhou (2013b) where a max-norm constrained maximum likelihood estimate is considered. This method allows more general non-uniform sampling schemes but still requires an upper bound on the max-norm of the unknown matrix. Here again, the rates of convergence obtained in Cai & Zhou (2013b) are slower than the rate of convergence of our estimator. Recently, Gunasekar et al. (2014) consider general exponential family distributions, which cover some distributions over finite sets. Their method, unlike our estimator, requires the knowledge of the "spikiness ratio" (usually unknown) and the uniform sampling scheme.

In the present paper, we consider a maximum likelihood estimator with nuclear-norm penalization. Our method allows us to consider general sampling scheme and only requires the knowledge of an upper bound on the maximum absolute value of the entries of the unknown matrix. All the previous works on this model also require the knowledge of this bound together with some additional (and more difficult to obtain) information on the unknown matrix.

The paper is organized as follows. In Section 2.2.1, the one-bit matrix completion is first discussed and our estimator is introduced. We establish upper bounds both on the Frobenius norm between the unknown true matrix and the proposed estimator and on the associated Kullback-Leibler divergence. In Section 2.2.2 lower bounds are established, showing that our upper bounds are minimax optimal up to logarithmic factors. Then, the one-bit matrix completion problem is extended to the case of a more general finite alphabet. In Section 2.3 an implementation based on the lifted coordinate descent algorithm recently introduced in Dudík et al. (2012) is proposed. A limited Monte Carlo experiment supporting our claims is then presented in Section 2.4.

Notations

For any integers $n, m_1, m_2 > 0$, $[n] := \{1, \dots, n\}$, $m_1 \vee m_2 := \max(m_1, m_2)$ and $m_1 \wedge m_2 := \min(m_1, m_2)$. We equip the set of $m_1 \times m_2$ matrices with real entries (denoted $\mathbb{R}^{m_1 \times m_2}$) with the scalar product $\langle X | X' \rangle := \mathrm{tr}(X^\top X')$. For a given matrix $X \in \mathbb{R}^{m_1 \times m_2}$ we write $\|X\|_\infty := \max_{i,j} |X_{i,j}|$ and for any $\rho \geq 1$, we denote its Schatten ρ-norm (see Bhatia (1997)) by

$$\|X\|_{\sigma,\rho} := \left(\sum_{i=1}^{m_1 \wedge m_2} \sigma_i^\rho(X) \right)^{1/\rho},$$

with $\sigma_i(X)$ the singular values of X ordered in decreasing order. The operator norm of X is $\|X\|_{\sigma,\infty} := \sigma_1(X)$. For any integer $q > 0$, we denote by $\mathbb{R}^{m_1 \times m_2 \times q}$ the set of $m_1 \times m_2 \times q$ (3-way) tensors. A tensor $\mathcal{X}$ is of the form $\mathcal{X} = (X^l)_{l=1}^q$ where $X^l \in \mathbb{R}^{m_1 \times m_2}$ for any $l \in [q]$. For any integer $p > 0$, a function $f : \mathbb{R}^q \to \mathcal{S}_p$ is called a p-link function, where $\mathcal{S}_p$ is the $p-$dimensional probability simplex. Given a p-link function f and $\mathcal{X}, \mathcal{X}' \in \mathbb{R}^{m_1 \times m_2 \times q}$, we define the squared Hellinger distance

$$\mathrm{d}_{\mathrm{H}}^2\left(f(\mathcal{X}), f(\mathcal{X}')\right) := \frac{1}{m_1 m_2} \sum_{k \in [m_1]} \sum_{k' \in [m_2]} \sum_{j \in [p]} \left[\left(\sqrt{f^j(\mathcal{X}_{k,k'})} - \sqrt{f^j(\mathcal{X}'_{k,k'})} \right)^2 \right],$$

where $\mathcal{X}_{k,k'}$ denotes the vector $(X^j_{k,k'})_{j=1}^q$. The Kullback-Leibler divergence is

$$\mathrm{KL}\left(f(\mathcal{X}), f(\mathcal{X}')\right) := \frac{1}{m_1 m_2} \sum_{k\in[m_1]} \sum_{k'\in[m_2]} \sum_{j\in[p]} \left[f^j(\mathcal{X}_{k,k'}) \log\left(\frac{f^j(\mathcal{X}_{k,k'})}{f^j(\mathcal{X}'_{k,k'})} \right) \right] .$$

For any tensor $\mathcal{X} \in \mathbb{R}^{m_1\times m_2\times q}$ we define $\mathrm{rk}(\mathcal{X}) := \max_{l\in[q]} \mathrm{rk}(X^l)$, where $\mathrm{rk}(X^l)$ is the rank of the matrix X^l and its sup-norm by $\|\mathcal{X}\|_\infty := \max_{l\in[q]} \|X^l\|_\infty$.

2.2 Main results

2.2.1 One-bit matrix completion

Assume that the observations follow a binomial distribution parametrized by a matrix $\bar{X} \in \mathbb{R}^{m_1\times m_2}$. Assume in addition that an $i.i.d.$ sequence of coefficients $(\omega_i)_{i=1}^n \in ([m_1]\times[m_2])^n$ is revealed and denote by Π their distribution. The observations associated to these coefficients are denoted by $(Y_i)_{i=1}^n \in \{1,2\}^n$ and distributed as follows

$$\mathbb{P}(Y_i = j) = f^j(\bar{X}_{\omega_i}), \quad j \in \{1,2\} , \tag{2.1}$$

where $f = (f^j)_{j=1}^2$ is a $2-$link function. For ease of notation, we often write $\bar{X}_i$ instead of $\bar{X}_{\omega_i}$. Denote by Φ_{Y} the (normalized) negative log-likelihood of the observations:

$$\Phi_{\mathrm{Y}}(X) = -\frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^2 \mathbb{1}_{\{Y_i=j\}} \log\left(f^j(X_i)\right) \right) . \tag{2.2}$$

Let $\gamma > 0$ be an upper bound of $\|\bar{X}\|_\infty$. We consider the following estimator of $\bar{X}$:

$$\hat{X} = \underset{X\in\mathbb{R}^{m_1\times m_2}, \|X\|_\infty\le\gamma}{\arg\min} \Phi(X) , \quad \text{where} \quad \Phi(X) = \Phi_{\mathrm{Y}}(X) + \lambda\|X\|_{\sigma,1} , \tag{2.3}$$

with $\lambda > 0$ being a regularization parameter. Consider the following assumptions.

H1. *The functions $x \mapsto -\ln(f^j(x))$, $j = 1,2$ are convex. In addition, There exist positive constants H_γ, L_γ and K_γ such that:*

$$H_\gamma \ge 2 \sup_{|x|\le\gamma} \left(|\log(f^1(x))| \vee |\log(f^2(x))| \right) , \tag{2.4}$$

$$L_\gamma \ge \max\left(\sup_{|x|\le\gamma} \frac{|(f^1)'(x)|}{f^1(x)}, \sup_{|x|\le\gamma} \frac{|(f^2)'(x)|}{f^2(x)} \right) , \tag{2.5}$$

$$K_\gamma = \inf_{|x|\le\gamma} g(x) , \quad \textit{where } g(x) = \frac{(f^1)'(x)^2}{8f^1(x)(1-f^1(x))} . \tag{2.6}$$

Remark 2.1. As shown in (Davenport et al., 2012, Lemma 2), K_γ satisfies

$$K_\gamma \le \inf_{\substack{x,y\in\mathbb{R}\\ |x|\le\gamma\\ |y|\le\gamma}} \left(\sum_{j=1}^2 \left(\sqrt{f^j(x)} - \sqrt{f^j(y)} \right)^2 / (x-y)^2 \right) . \tag{2.7}$$

Our framework allows a general distribution Π. We assume that Π satisfies the following assumptions introduced in Klopp (2014) in the classical setting of unquantized matrix completion:

H2. *There exists a constant $\mu > 0$ such that, for any $m_1 > 0$ and $m_2 > 0$*

$$\min_{k\in[m_1],k'\in[m_2]} \pi_{k,k'} \geq \mu/(m_1 m_2) \ , \quad \text{where } \pi_{k,k'} = \mathbb{P}(\omega_1 = (k,k')) \ . \tag{2.8}$$

Denote by $R_k = \sum_{k'=1}^{m_2} \pi_{k,k'}$ and $C_{k'} = \sum_{k=1}^{m_1} \pi_{k,k'}$ the probability of revealing a coefficient from row k and column k', respectively.

H3. *There exists a constant $\nu \geq 1$ such that, for all m_1, m_2,*

$$\max_{k,l}(R_k, C_l) \leq \frac{\nu}{m_1 \wedge m_2} \ ,$$

The first assumption ensures that every coefficient has a nonzero probability of being observed, whereas the second assumption requires that no column nor row is sampled with too high probability (see also Foygel et al. (2011); Klopp (2014) for more details on these conditions). For instance, the uniform distribution yields $\mu = \nu = 1$. Define

$$d = m_1 + m_2 \ , \quad M = m_1 \vee m_2, \quad m = m_1 \wedge m_2 \ . \tag{2.9}$$

Theorem 2.2. *Assume H1, H2, H3 and that $\|\bar{X}\|_\infty \leq \gamma$. Assume in addition that $n \geq 2m\log(d)/(9\nu)$. Take*

$$\lambda = 6L_\gamma \sqrt{\frac{2\nu \log(d)}{mn}} \ .$$

Then, with probability at least $1 - 3d^{-1}$ the Kullback-Leibler divergence is bounded by

$$\mathrm{KL}\left(f(\bar{X}), f(\hat{X})\right) \leq \mu \max\left(\bar{c}\mu\nu \frac{L_\gamma^2 \operatorname{rk}(\bar{X})}{K_\gamma} \frac{M\log(d)}{n}, \mathrm{e}H_\gamma \sqrt{\frac{\log(d)}{n}}\right) \ ,$$

with $\bar{c}$ a universal constant whose value is specified in the proof.

Proof. See Section 2.5.1. □

This result immediately gives an upper bound on the estimation error of $\hat{X}$, measured in Frobenius norm:

Corollary 2.3. *Under the same assumptions and notations of Theorem 2.2 we have with probability at least $1 - 3d^{-1}$*

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \mu \max\left(\bar{c}\mu\nu \frac{L_\gamma^2 \operatorname{rk}(\bar{X})}{K_\gamma^2} \frac{M\log(d)}{n}, \frac{\mathrm{e}H_\gamma}{K_\gamma}\sqrt{\frac{\log(d)}{n}}\right) \ .$$

Proof. Using Theorem 2.11 and Theorem 2.2, the result follows. □

Remark 2.4. Note that, up to the factor L_γ^2/K_γ^2, the rate of convergence given by Theorem 2.3, is the same as in the case of usual unquantized matrix completion, see, for example, Klopp (2014) and Koltchinskii et al. (2011). For this usual matrix completion setting, it has been shown in (Koltchinskii et al., 2011, Theorem 3) that this rate is minimax optimal up to a

logarithmic factor. Let us compare this rate of convergence with those obtained in previous works on 1-bit matrix completion. In Davenport et al. (2012), the parameter $\bar{X}$ is estimated by minimizing the negative log-likelihood under the constraints $\|X\|_\infty \leq \gamma$ and $\|X\|_{\sigma,1} \leq \gamma\sqrt{rm_1m_2}$ for some $r > 0$. Under the assumption that $\mathrm{rk}(\bar{X}) \leq r$, they could prove that

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1m_2} \leq C_\gamma \sqrt{\frac{rd}{n}} \,,$$

where C_γ is a constant depending on γ (see (Davenport et al., 2012, Theorem 1)). This rate of convergence is slower than the rate of convergence given by Theorem 2.3. Cai & Zhou (2013b) studied a max-norm constrained maximum likelihood estimate and obtain a rate of convergence similar to Davenport et al. (2012). In Gunasekar et al. (2014), matrix completion was considered for a likelihood belonging to the exponential family. Note, for instance, that the logit distribution belongs to such a family. The following upper bound on the estimation error is provided (see (Gunasekar et al., 2014, Theorem 1))

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1m_2} \leq C_\gamma \left(\alpha_*^2 \frac{\mathrm{rk}(\bar{X}) M \log(M)}{n} \right). \tag{2.10}$$

Comparing with Theorem 2.3, (2.10) contains an additional term α_*^2 where α_* is an upper bound of $\sqrt{m_1m_2}\|\bar{X}\|_\infty$.

2.2.2 Minimax lower bounds for one-bit matrix completion

Theorem 2.3 insures that our estimator achieves certain Frobenius norm errors. We now discuss the extent to which this result is optimal. A classical way to address this question is by determining minimax rates of convergence.

For any integer $0 \leq r \leq \min(m_1, m_2)$ and any $\gamma > 0$, we consider the following family of matrices

$$\mathcal{F}(r, \gamma) = \left\{ \bar{X} \in \mathbb{R}^{m_1 \times m_2} : \mathrm{rank}(\bar{X}) \leq r, \, \|\bar{X}\|_\infty \leq \gamma \right\}.$$

We will denote by $\inf_{\hat{X}}$ the infimum over all estimators $\hat{X}$ that are functions of the data $(\omega_i, Y_i)_{i=1}^n$. For any $X \in \mathbb{R}^{m_1 \times m_2}$, let $\mathbb{P}_X$ denote the probability distribution of the observations $(\omega_i, Y_i)_{i=1}^n$ for a given $2-$link function f and sampling distribution Π. We establish a lower bound under an additional assumption on the function f^1:

H4. *$(f^1)'$ is decreasing on $\mathbb{R}_+$ and $K_\gamma = g(\gamma)$ where g and K_γ are defined in (2.6).*

In particular, H4 is satisfied in the case of logit or probit models. The following theorem establishes a lower bound on the minimax risk in squared Frobenius norm:

Theorem 2.5. *Assume H4. Let $\alpha \in (0, 1/8)$ Then there exists a constant $c > 0$ such that, for all $m_1, m_2 \geq 2$, $1 \leq r \leq m$, and $\gamma > 0$,*

$$\inf_{\hat{X}} \sup_{\bar{X} \in \mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}} \left(\frac{\|\hat{X} - \bar{X}\|_2^2}{m_1m_2} > c \min\left\{ \gamma^2, \frac{Mr}{n K_0} \right\} \right) \geq \delta(\alpha, M) \,,$$

where

$$\delta(\alpha, M) = \frac{1}{1 + 2^{-rM/16}} \left(1 - 2\alpha - \frac{1}{2}\sqrt{\frac{\alpha}{\log(2)(rM)}} \right). \tag{2.11}$$

Proof. See Section 2.5.3. □

Note that the lower bound given by Theorem 2.5 is proportional to the rank multiplied by the maximum dimension of $\bar{X}$ and inversely proportional the sample size n. Therefore the lower bound matches the upper bound given by Theorem 2.3 up to a constant and a logarithmic factor. The lower bound does not capture the dependance on γ, note however that the upper and lower bound only differ by a factor L_γ^2 / K_γ.

2.2.3 Extension to multi-class problems

Let us now consider a more general setting where the observations follow a distribution over a finite set $\{1, \dots, p\}$, parameterized by a tensor $\bar{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$. The distribution of the observations $(Y_i)_{i=1}^n \in [p]^n$ is

$$\mathbb{P}(Y_i = j) = f^j(\bar{\mathcal{X}}_{\omega_i}), \quad j \in [p] \ ,$$

where $f = (f^j)_{j=1}^p$ is now a p-link function and $\bar{\mathcal{X}}_{\omega_i}$ denotes the vector $(\bar{X}^l_{\omega_i})_{l=1}^q$. The negative log-likelihood of the observations is now given by:

$$\Phi_{\mathrm{Y}}(\mathcal{X}) = -\frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \mathbb{1}_{\{Y_i = j\}} \log \left(f^j(\mathcal{X}_i) \right) \right) \ . \tag{2.12}$$

where we use the notation $\mathcal{X}_i = \mathcal{X}_{\omega_i}$. Our proposed the estimator is defined as:

$$\hat{\mathcal{X}} = \underset{\substack{\mathcal{X} \in \mathbb{R}^{m_1 \times m_2 \times q} \\ \|\mathcal{X}\|_\infty \leq \gamma}}{\arg\min} \Phi(\mathcal{X}) \ , \quad \text{where} \quad \Phi(\mathcal{X}) = \Phi_{\mathrm{Y}}(\mathcal{X}) + \lambda \sum_{j=1}^q \|X^j\|_{\sigma,1} \ , \tag{2.13}$$

In order to extend the results of the previous sections we make an additional assumption which allows to split the log-likelihood as a sum.

H5. *There exist functions $(g_l^j)_{(l,j) \in [p] \times [q]}$ such that the p-link function f can be factorized as follows*

$$f^j(x_1, \dots, x_q) = \prod_{l=1}^q g_l^j(x_l) \quad \textit{for } j \in [p] \ .$$

The model considered above covers many finite distributions including among others logistic binomial (see Section 2.2.1) and conditional logistic multinomial (see Section 2.3).

Assumptions on constants depending on the link function are extended by

H6. *There exist positive constant H_γ, L_γ and K_γ such that:*

$$H_\gamma \geq \max_{(j,l) \in [p] \times [q]} \sup_{|x| \leq \gamma} 2 |\log(g_l^j(x))| \ , \tag{2.14}$$

$$L_\gamma \geq \max_{(j,l) \in [p] \times [q]} \sup_{|x| \leq \gamma} \left| \frac{(g_l^j)'(x)}{g_l^j(x)} \right| \ , \tag{2.15}$$

$$K_\gamma \leq \inf_{\substack{x,y \in \mathbb{R}^q \\ \|x\|_\infty \leq \gamma \\ \|y\|_\infty \leq \gamma}} \left(\sum_{j=1}^p \left(\sqrt{f^j(x)} - \sqrt{f^j(y)} \right)^2 / \|x - y\|_2^2 \right) \ . \tag{2.16}$$

For any tensor $\mathcal{X} \in \mathbb{R}^{m_1 \times m_2 \times q}$, we write $\bar{\Sigma} := \nabla \Phi_Y(\bar{\mathcal{X}}) \in \mathbb{R}^{m_1 \times m_2 \times q}$. We also define the sequence of matrices $(E_i)_{i=1}^n$ associated to the revealed coefficients $(\omega_i)_{i=1}^n$ by $E_i := e_{k_i}(e'_{l_i})^\top$ where $(k_i, l_i) = \omega_i$ and with $(e_k)_{k=1}^{m_1}$ (*resp.* $((e'_l)_{l=1}^{m_2})$ being the canonical basis of $\mathbb{R}^{m_1}$ (*resp.* $\mathbb{R}^{m_2}$). Furthermore, if $(\varepsilon_i)_{1 \le i \le n}$ is a Rademacher sequence independent from $(\omega_i)_{i=1}^n$ and $(Y_i)_{1 \le i \le n}$ we define the matrix Σ_R as follow

$$\Sigma_R := \frac{1}{n} \sum_{i=1}^n \varepsilon_i E_i \ .$$

We can now state the main results of this paper.

Theorem 2.6. *Assume H2, H5 and H6 hold,* $\lambda > 2 \max_{l \in [q]} \|\bar{\Sigma}^l\|_{\sigma,\infty}$ *and* $\left\|\bar{\mathcal{X}}\right\|_\infty \le \gamma$. *Then, with probability at least* $1 - 2d^{-1}$, *the Kullback-Leibler divergence is bounded by*

$$\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\hat{\mathcal{X}})\right) \le \mu \max\left(4\mu \frac{m_1 m_2 \, \mathrm{rk}(\bar{\mathcal{X}})}{K_\gamma} \left(\lambda^2 + 256 \mathrm{e}(q L_\gamma \mathbb{E}\left\|\Sigma_R\right\|_{\sigma,\infty})^2\right), \mathrm{e} H_\gamma \sqrt{\frac{\log(d)}{n}}\right) \ .$$

with d *defined in* (2.9).

Proof. See Section 2.5.1. □

Note that the lower bound of λ is stochastic and the expectation $\mathbb{E}\|\Sigma_R\|_{\sigma,\infty}$ is unknown. However, these quantities can be controlled using H3.

Theorem 2.7. *Assume H2, H3, H5 and H6 hold and that* $\|\bar{\mathcal{X}}\|_\infty \le \gamma$. *Assume in addition that* $n \ge 2m \log(d)/(9\nu)$. *Take*

$$\lambda = 6 L_\gamma \sqrt{\frac{2\nu \log(d)}{mn}} \ .$$

Then, with probability at least $1 - (2+q)d^{-1}$, *the Kullback-Leibler divergence is bounded by*

$$\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\hat{\mathcal{X}})\right) \quad \le \quad \mu \max\left(\bar{c}\mu\nu \frac{q^2 L_\gamma^2 \, \mathrm{rk}(\bar{\mathcal{X}})}{K_\gamma} \frac{M \log(d)}{n}, \mathrm{e} H_\gamma \sqrt{\frac{\log(d)}{n}}\right) \quad ,$$

with $\bar{c}$ *a universal constant ,* d, m *and* M *defined in* (2.9).

Proof. See Section 2.5.2. □

2.3 Implementation

In this section an implementation for the following p-class link function is given:

$$f^j(x^1, \ldots, x^{p-1}) = \begin{cases} \exp(x^j) \left(\prod_{l=1}^j (1 + \exp(x^l))\right)^{-1} & \text{if } j \in [p-1] \ , \\ \left(\prod_{l=1}^{p-1} (1 + \exp(x^l))\right)^{-1} & \text{if } j = p \ . \end{cases}$$

This p-class link function boils down to parameterizing the distribution of the observation as follows:

$$\mathbb{P}(Y_i = 1) = \frac{\exp(\bar{X}_i^1)}{1 + \exp(\bar{X}_i^1)} ,$$
$$\mathbb{P}(Y_i = j | Y_i > j - 1) = \frac{\exp(\bar{X}_i^j)}{1 + \exp(\bar{X}_i^j)} \quad \text{for } j \in \{2, \dots, p-1\} .$$

Assumption H5 is satisfied and the problem (2.13) is separable *w.r.t.* each matrix X^l. Following Davenport et al. (2012), we solve (2.13) without taking into account the constraint γ; as reported in Davenport et al. (2012) and confirmed by our experiments, the impact of this projection is negligible, whereas it increases significantly the computation burden.

Because the problem is separable, it suffices to solve in parallel each sub-problem

$$\hat{X}^l = \underset{X \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \ \Phi_\lambda^l(X) , \quad \text{where} \quad \Phi_\lambda^l(X) = \Phi^l(X) + \lambda \|X\|_{\sigma,1} . \tag{2.17}$$

This can be achieved by using the coordinate gradient descent algorithm introduced by Dudík et al. (2012). To describe the algorithm, consider first the set of normalized rank one matrices

$$\mathcal{M} := \left\{ M \in \mathbb{R}^{m_1 \times m_2} | M = uv^\top \mid \|u\|_2 = \|v\|_2 = 1, \right\} .$$

Define Θ the linear space of real-valued functions on $\mathcal{M}$ with finite support, *i.e.,* any $\theta \in \Theta$ satisfies $\theta(M) = 0$ except for a finite number of $M \in \mathcal{M}$. This space is equipped with the ℓ^1-norm $\|\theta\|_1 = \sum_{M \in \mathcal{M}} |\theta(M)|$. Define by Θ_+ the positive orthant, *i.e.,* the cone of functions $\theta \in \Theta$ such that $\theta(M) \geq 0$ for all $M \in \mathcal{M}$. Any matrix $X \in \mathbb{R}^{m_1 \times m_2}$ can be associated to an element $\theta \in \Theta_+$ satisfying

$$X = \sum_{M \in \mathcal{M}} \theta(M) M . \tag{2.18}$$

Such function is not unique. Consider an SVD of X *i.e.,* $X = \sum_{i=1}^m \lambda_i u_i v_i^\top$, where $(\lambda_i)_{i=1}^m$ are the singular values and $(u_i)_{i=1}^m$, $(v_i)_{i=1}^m$ are left and right singular vectors, then $\theta_X = \sum_{i=1}^m \lambda_i \delta_{u_i v_i^\top}$ satisfies (2.18), with $\delta_M \in \Theta$ is the function on $\mathcal{M}$ satisfying $\delta_M(M) = 1$ and $\delta_M(M') = 0$ if $M' \neq M$. As seen below, the function θ_X plays a key role.

Conversely, for any $\theta \in \Theta_+$, define

$$W : \theta \to W_\theta := \sum_{M \in \mathcal{M}} \theta(M) M .$$

and the auxiliary objective function:

$$\tilde{\Phi}_\lambda^l : \theta \to \tilde{\Phi}_\lambda^l(\theta) := \lambda \sum_{M \in \mathcal{M}} \theta(M) + \Phi^l(W_\theta) . \tag{2.19}$$

The triangular inequality implies that for all $\theta \in \Theta_+$,

$$\|W_\theta\|_{\sigma,1} \leq \|\theta\|_1 .$$

For $\theta \in \Theta$ we denote by $\mathrm{supp}(\theta)$ the support of θ *i.e.,* the subset of $\mathcal{M}$ such that $\theta(M) \neq 0 \iff M \in \mathrm{supp}(\theta)$. If for any $M, M' \in \mathrm{supp}(\theta)$, $M \neq M'$, $\langle M, M' \rangle = 1$, then $\|\theta\|_1 = \|W_\theta\|_{\sigma,1}$. Indeed in such case $\sum_{M \in \mathcal{M}} \theta(M) M$ defines a SVD of W_θ. Therefore the minimization of (2.19) is actually equivalent to the minimization of (2.17); see (Dudík et al.,

Parameter Size	1000×1000	3000×3000	10000×10000
Observations	$100 \cdot 10^3$	$1 \cdot 10^6$	$10 \cdot 10^6$
Execution Time (s.)	4.5	52	730

TABLE 2.1: Execution time of the proposed algorithm for the binary case.

2012, Theorem 3.2). The minimization (2.19) can be implemented using a coordinate gradient descent algorithm which updates at each iteration the nonnegative finite support function θ.

Algorithm 8 Lifted coordinate gradient descent

Initialization: initial parameter θ_0, precision ϵ
Loop:
  Compute the top singular vector pair of $-\nabla \Phi^l(W_{\theta_k})$: u_k, v_k
  $g_k \leftarrow \lambda + \langle \nabla \Phi^l(W_{\theta_k}), u_k v_k^\top \rangle$
  if $g_k \leq -\epsilon/2$ **then**
    $\beta_k \leftarrow \arg\min_{b \in \mathbb{R}_+} \tilde{\Phi}^l_\lambda \left(\theta + b\delta_{u_k v_k^\top} \right)$
    $\theta_{k+1} \leftarrow \theta_k + \beta_k \delta_{u_k v_k^\top}$,
  else
    $g_k^{\max} \leftarrow \max_{M \in \mathrm{supp}(\theta_k)} |\lambda + \langle \nabla \Phi^l(W_{\theta_k}), M \rangle|$
    if $g_k^{\max} \leq \epsilon$ **then**
      Break
    else
      $\theta_{k+1} \leftarrow \arg\min_{\theta' \in \Theta_+, \mathrm{supp}(\theta') \subset \mathrm{supp}(\theta_k)} \tilde{\Phi}^l_\lambda(\theta')$
    end if
  end if

The algorithm is summarized in Algorithm 8. Compared to the Soft-Impute Mazumder et al. (2010) or the SVT Cai et al. (2010) algorithms, this algorithm does not require the computation of a full SVD at each step of the main loop of an iterative (proximal) algorithm (recall that the proximal operator associated to the nuclear norm is the soft-thresholding operator of the singular values). The proposed algorithm requires only to compute the largest singular values and associated singular vectors.

Another interest of this algorithm is that it only requires to evaluate the coordinate of the gradient for the entries which have been actually observed. It is therefore memory efficient when the number of observations is smaller than the total number of coefficients $m_1 m_2$, which is the typical setting in which matrix completion is used. Moreover, we use Arnoldi iterations to compute the top singular values and vector pairs (see (Golub & van Loan, 2013, Section 10.5) for instance) which allows us to take full advantage of sparse structures, the minimizations in the inner loop are carried out using the L-BFGS-B algorithm. Table 2.1 provides the execution time one-bit matrix completion (on a 3.07Ghz w3550 Xeon CPU with RAM 1.66 Go, Cache 8 Mo, C implementation).

2.4 Numerical Experiments

We have performed numerical experiments on both simulated and real data provided by the MovieLens project (`http://grouplens.org`). Both the one-bit matrix completion - $p = 2$, $q = 1$ - and the extended multi-class setting -$p = 5$, $q = 4$ - are considered; comparisons are

also provided with the classical Gaussian matrix completion algorithm to assess the potential gain achieved by explicitly taking into account the facts that the observations belong to a finite alphabet. Only a limited part of the experiments are reported in this article; a more extensive assessment can be obtained upon authors request.

For each matrix $\bar{X}^l$ we sampled uniformly five unitary (for the Euclidean norm) vector pairs $(u_k^l, v_k^l)_{k=1}^5$. The matrix $\bar{X}^l$ is then defined as

$$\bar{X}^l = \Gamma\sqrt{m_1 m_2}\sum_{k=1}^{5}\alpha_k u_k^l (v_k^l)^\top + \eta^l \mathbf{I}_{m_1\times m_2},$$

with $(\alpha_1,\ldots,\alpha_5) = (2, 1, 0.5, 0.25, 0.1)$, Γ a scaling factor and $\mathbf{I}_{m_1\times m_2}$ the $m_1 \times m_2$ matrix of ones. The term η_l has been fixed so that each class has the same average probability *i.e.,* $f^j((\mathbb{E}[\bar{X}^l])_{l=1}^{p-1}) = 1/p$ for $j \in [p]$. Note that the factor $\sqrt{m_1 m_2}$ implies that the variance of $\bar{X}^l$ coefficients does not depend on m_1 and m_2. The sizes investigated are $(m_1, m_2) \in \{(500, 300), (1000, 600)\}$.

The observations are sampled to the conditional multinomial logistic model introduced in Section 2.3. For comparison purposes we have also computed $\hat{X}^{\mathcal{N}}$, the classical Gaussian version (*i.e.,* using a squared Frobenius norm in (2.13)). Contrary to the logit version, the Gaussian matrix completion does not directly recover the distribution of the observations $(Y_i)_{i=1}^n$. However, we can estimate $\mathbb{P}(Y_i = j)$ by the following quantity:

$$F_{\mathcal{N}(0,1)}(p_{j+1}) - F_{\mathcal{N}(0,1)}(p_j) \text{ with } p_j = \begin{cases} 0 & \text{if } j = 1\ , \\ \frac{j-0.5-\hat{X}_i^{\mathcal{N}}}{\hat{\sigma}} & \text{if } 0 < j < p \\ 1 & \text{if } j = p\ , \end{cases}$$

where $F_{\mathcal{N}(0,1)}$ is the cdf of a zero-mean standard Gaussian random variable.

The choice of the regularization parameter λ has been solved for all methods by performing 5-fold cross-validation on a geometric grid of size $0.6\log(n)$ (note that the estimators are null for λ greater than $\|\nabla\,\Phi_{\mathrm{Y}}(0)\|_{\sigma,\infty}$).

As evidenced in Figure 2.1, the Kullback-Leibler divergence for the logistic estimator is significantly lower than for the Gaussian estimator, for both the $p = 2$ and $p = 5$ cases. This was expected because the Gaussian model assume implicitly symmetric distributions with the same variance for all the ratings, These assumptions are of course avoided by the logistic modem.

Regarding the prediction error, Table 2.2 and Table 2.3 summarize the results obtained for a 1000×600 matrix. The logistic model outperforms the Gaussian model (slightly for $p = 2$ and significantly for $p = 5$).

Number of observations	$10 \cdot 10^3$	$50 \cdot 10^3$	$250 \cdot 10^3$	$500 \cdot 10^3$
Gaussian prediction error	0.50	0.38	0.32	0.32
Logistic prediction error	0.46	0.33	0.31	0.31

TABLE 2.2: Prediction errors for a binomial (2 classes) underlying model, for a 1000×600 matrix.

We have also run the same estimators on the MovieLens $100k$ dataset. In this case, the Kullback-Leibler divergence cannot be computed. Therefore, to assess the prediction errors, we randomly select 20% of the entries as a test set, and the remaining entries are split between a training set (80%) and a validation set (20%).

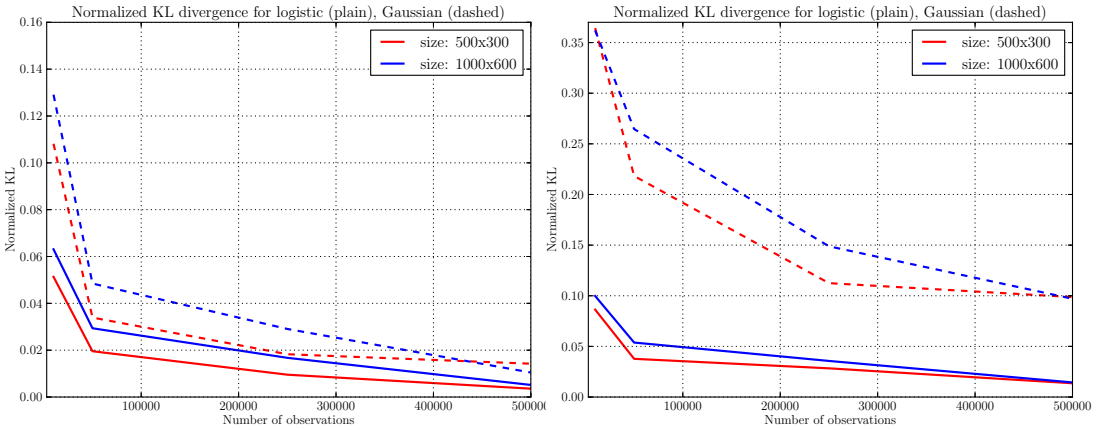

FIGURE 2.1: Kullback-Leibler divergence between the estimated and the true model for different matrices sizes and sampling fraction, normalized by number of classes. Right figure: binomial and the Gaussian models ; left figure: multinomial with five classes and Gaussian model.

Number of observations	$10 \cdot 10^3$	$50 \cdot 10^3$	$250 \cdot 10^3$	$500 \cdot 10^3$
Gaussian prediction error	0.75	0.75	0.72	0.71
Logistic prediction error	0.75	0.67	0.58	0.57

TABLE 2.3: Prediction Error for a multinomial (5 classes) distribution against a 1000×600 matrix.

For this dataset, ratings range from 1 to 5. To consider the benefit of a binomial model, we have tested each rating against the others (*e.g.,* ratings 5 are set to 0 and all others are set to 1).

These results are summarized in Table 2.4. For the multinomial case, we find a prediction error of 0.59 for the logistic model against a 0.63 for the Gaussian one.

Rating against the others	1	2	3	4	5
Gaussian prediction error	0.12	0.20	0.39	0.46	0.30
Logistic prediction error	0.06	0.11	0.27	0.34	0.20

TABLE 2.4: Binomial prediction error when performing one versus the others procedure on the MovieLens $100k$ dataset.

2.5 Proofs of main results

2.5.1 Proof of Theorem 2.2 and Theorem 2.6

Proof. Since Theorem 2.2 is an application of Theorem 2.6 for $p = 2$ and $q = 1$ it suffices to prove Theorem 2.6.

We consider a tensor $\mathcal{X}$ which satisfies $\Phi(\mathcal{X}) \leq \Phi(\bar{\mathcal{X}})$, (*e.g.,* $\mathcal{X} = \hat{\mathcal{X}}$). We get from Theorem 2.8

$$\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) \leq \lambda \sqrt{\bar{r}} \sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}\,, \tag{2.20}$$

where

$$\bar{r} = \frac{2m_1 m_2 \,\mathrm{rk}(\tilde{\mathcal{X}})}{K_\gamma}. \tag{2.21}$$

Let us define

$$\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) := \mathbb{E}\left[\left(\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}})\right)\right] , \tag{2.22}$$

where the expectation is taken both over the $(E_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$. As stated in Theorem 2.13, H2 implies $\mu\,\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \geq \mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)$. We now need to control the left hand side of (2.20) uniformly over X with high probability. Since we assume $\lambda > 2\max_{l\in[q]} \|\bar{\Sigma}^l\|_{\sigma,\infty}$ applying Theorem 2.12 (2.30) and then Theorem 2.13 yields

$$\sum_{l=1}^{q} \|X^l - \bar{X}^l\|_{\sigma,1} \leq 4\sqrt{\bar{r}}\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)} \leq 4\sqrt{\mu\bar{r}}\sqrt{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)} , \tag{2.23}$$

Consequently, if we define $\mathcal{C}(r)$ as

$$\mathcal{C}(r) := \left\{ \mathcal{X} \in \mathbb{R}^{m_1\times m_2\times q} : \ \sum_{l=1}^{q} \|X^l - \bar{X}^l\|_{\sigma,1} \leq \sqrt{r\,\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)} \right\} ,$$

we need to control $(\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}))$ for $\mathcal{X} \in \mathcal{C}(16\mu\bar{r})$. We have to ensure that $\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)$ is greater than a given threshold $\beta > 0$ and therefore we define the following set

$$\mathcal{C}_\beta(r) = \left\{\mathcal{X} \in \mathcal{C}(r),\ \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \geq \beta\right\} . \tag{2.24}$$

We then consider the two following cases.
Case 1. If $\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) > \beta$, (2.23) gives $X \in \mathcal{C}_\beta(16\mu\bar{r})$. Plugging Theorem 2.14 in (2.20) with $\beta = 2M_\gamma\sqrt{\log(d)}/(\eta\sqrt{n\log(\alpha)})$, $\alpha = e$ and $\eta = 1/(4\alpha)$ then it holds with probability at least $1 - 2d^{-1}/(1-d^{-1}) \geq 1 - 2/d$

$$\frac{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2} - \epsilon(16\mu\bar{r}, \alpha, \eta) \leq \lambda\sqrt{\bar{r}}\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)} ,$$

where ϵ is defined in Theorem 2.14. Recalling Theorem 2.13 we get

$$\frac{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2\mu} - \lambda\sqrt{\bar{r}}\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}, f(\mathcal{X})\right)} - \epsilon(16\mu\bar{r}, \alpha, \eta) \leq 0 .$$

An analysis of this second order polynomial and the relation $\epsilon(16\mu\bar{r}, \alpha, \eta)/\mu = \epsilon(16\bar{r}, \alpha, \eta)$ lead to

$$\sqrt{\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)} \leq \mu\left(\lambda\sqrt{\bar{r}} + \sqrt{\lambda^2\bar{r} + 2\epsilon(16\bar{r}, \alpha, \eta)}\right) . \tag{2.25}$$

Applying the inequality $(a+b)^2 \leq 2(a^2+b^2)$ gives the bound of Theorem 2.6.
Case 2. If $\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \leq \beta$ then Theorem 2.13 yields

$$\mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \leq \mu\beta . \tag{2.26}$$

Combining (2.25) and (2.26) concludes the proof. □

For $X \in \mathbb{R}^{m_1\times m_2}$, denote by $\mathcal{S}_1(X) \subset \mathbb{R}^{m_1}$ (*resp.* $\mathcal{S}_2(X) \subset \mathbb{R}^{m_2}$) the linear spans generated by left (*resp.* right) singular vectors of X. $P_{\mathcal{S}_1^\perp(X)}$ (*resp.* $P_{\mathcal{S}_2^\perp(X)}$) denotes the orthogonal projections on $\mathcal{S}_1^\perp(X)$ (*resp.* $\mathcal{S}_2^\perp(X)$). We then define the following orthogonal projections on $\mathbb{R}^{m_1\times m_2}$

$$\Pi_{\mathcal{C}_X^\perp} : \tilde{X} \mapsto P_{\mathcal{S}_1^\perp(X)}\tilde{X}P_{\mathcal{S}_2^\perp(X)} \text{ and } \Pi_{\mathcal{C}X} : \tilde{X} \mapsto \tilde{X} - \Pi_{\mathcal{C}_X^\perp}(\tilde{X}) .$$

Lemma 2.8. *Let $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ satisfying $\Phi(\mathcal{X}) \leq \Phi(\tilde{\mathcal{X}})$, then*

$$\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) \leq \lambda \bar{r}^{1/2} \sqrt{\mathrm{KL}\left(f(\tilde{\mathcal{X}}), f(\mathcal{X})\right)} \,,$$

where $\bar{r}$ is defined in (2.21).

Proof. Since $\Phi(\mathcal{X}) \leq \Phi(\tilde{\mathcal{X}})$, we obtain

$$\begin{aligned}\Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) \leq & \lambda \sum_{l=1}^{q} (\|\tilde{X}^l\|_{\sigma,1} - \|X^l\|_{\sigma,1}) \leq \lambda \sum_{l=1}^{q} \|\,\Pi_{\mathcal{C}_{\tilde{X}^l}}(X - \tilde{X}^l)\|_{\sigma,1} \,, \\ \leq & \lambda \sqrt{2\,\mathrm{rk}(\tilde{\mathcal{X}})} \left(\sum_{l=1}^{q} \|X - \tilde{X}\|_{\sigma,2} \right) \,,\end{aligned}$$

where we have used Theorem 2.9-(ii) and (iii) and for the last two lines and the definition of K_γ and Theorem 2.10 to get the result. □

Lemma 2.9. *For any pair of matrices X, $\tilde{X} \in \mathbb{R}^{m_1 \times m_2}$ we have*

(i) $\|X + \Pi_{\mathcal{C}_X^\perp}(\tilde{X})\|_{\sigma,1} = \|X\|_{\sigma,1} + \|\,\Pi_{\mathcal{C}_X^\perp}(\tilde{X})\|_{\sigma,1} \,,$

(ii) $\|\,\Pi_{\mathcal{C}_X}(\tilde{X})\|_{\sigma,1} \leq \sqrt{2\,\mathrm{rk}(X)}\|\tilde{X}\|_{\sigma,2} \,,$

(iii) $\|X\|_{\sigma,1} - \|\tilde{X}\|_{\sigma,1} \leq \|\,\Pi_{\mathcal{C}_X}(\tilde{X} - X)\|_{\sigma,1} \,.$

Proof. If $A, B \in \mathbb{R}^{m_1 \times m_2}$ are two matrices satisfying $\mathcal{S}_i(A) \perp \mathcal{S}_i(B)$, $i = 1, 2$, then $\|A + B\|_{\sigma,1} = \|A\|_{\sigma,1} + \|B\|_{\sigma,1}$. Applying this identity with $A = X$ and $B = \Pi_{\mathcal{C}_X^\perp}(\tilde{X})$, we obtain

$$\|X + \Pi_{\mathcal{C}_X^\perp}(\tilde{X})\|_{\sigma,1} = \|X\|_{\sigma,1} + \|\,\Pi_{\mathcal{C}_X^\perp}(\tilde{X})\|_{\sigma,1} \,,$$

showing (i).

It follows from the definition that $\Pi_{\mathcal{C}_X}(\tilde{X}) = P_{\mathcal{S}_1(X)}\tilde{X}P_{\mathcal{S}_2^\perp(X)} + \tilde{X}P_{\mathcal{S}_2(X)}$. Note that $\Pi_{\mathcal{C}_X}$ is an orthogonal projector on $\mathbb{R}^{m_1 \times m_2}$ equipped with the euclidean product $\langle \cdot, \cdot \rangle$. On the other hand, the Cauchy-Schwarz inequality implies that for any matrix C, $\|C\|_{\sigma,1} \leq \sqrt{\mathrm{rk}(C)}\|C\|_{\sigma,2}$. Consequently (ii) follows from

$$\|\,\Pi_{\mathcal{C}_X}(\tilde{X})\|_{\sigma,1} \leq \sqrt{2\,\mathrm{rk}(X)}\|\,\Pi_{\mathcal{C}_X}(\tilde{X})\|_{\sigma,2} \leq \sqrt{2\,\mathrm{rk}(X)}\|\tilde{X}\|_{\sigma,2} \,.$$

Finally, since $\tilde{X} = X + \Pi_{\mathcal{C}_X^\perp}(\tilde{X} - X) + \Pi_{\mathcal{C}_X}(\tilde{X} - X)$ we have

$$\begin{aligned}\|\tilde{X}\|_{\sigma,1} &\geq \|X + \Pi_{\mathcal{C}_X^\perp}(\tilde{X} - X)\|_{\sigma,1} - \|\,\Pi_{\mathcal{C}_X}(\tilde{X} - X)\|_{\sigma,1} \,, \\ &= \|X\|_{\sigma,1} + \|\,\Pi_{\mathcal{C}_X^\perp}(\tilde{X} - X)\|_{\sigma,1} - \|\,\Pi_{\mathcal{C}_X}(\tilde{X} - X)\|_{\sigma,1} \,,\end{aligned}$$

leading to (iii). □

Lemma 2.10. *For any tensor $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ and p-link function f it holds:*

$$\mathrm{d}_{\mathrm{H}}^2\left(f(\mathcal{X}), f(\tilde{\mathcal{X}})\right) \leq \mathrm{KL}\left(f(\mathcal{X}), f(\tilde{\mathcal{X}})\right)$$

Proof. See (Tsybakov, 2009, Lemma 4.2) □

Lemma 2.11. *For any $p, q > 0$ and p-link function f and any $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ satisfying $\|\mathcal{X}\|_\infty \leq \gamma$ and $\|\tilde{\mathcal{X}}\|_\infty \leq \gamma$, we get:*

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,2}^2 \leq \frac{m_1 m_2}{K_\gamma} \mathrm{d}_{\mathrm{H}}^2 \left(f(\mathcal{X}), f(\tilde{\mathcal{X}}\right) \leq \frac{m_1 m_2}{K_\gamma} \mathrm{KL}\left(f(\mathcal{X}), f(\tilde{\mathcal{X}}\right) .$$

Proof. For $p = 2$ and $q = 1$, it is a consequence of Theorem 2.1 and Theorem 2.10. Otherwise, the proof follows from the definition (2.16) of K_γ and Theorem 2.10. □

Lemma 2.12. *Let $\mathcal{X}, \tilde{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times q}$ satisfying $\|\mathcal{X}\|_\infty \leq \gamma$ and $\|\tilde{\mathcal{X}}\|_\infty \leq \gamma$. Assume that $\lambda > 2 \max_{l \in [q]} \|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty}$ and $\Phi(X) \leq \Phi(\tilde{X})$. Then*

$$\sum_{l=1}^{q} \| \Pi_{\mathcal{C}_{\tilde{X}^l}^{\perp}}(X^l - \tilde{X}^l)\|_{\sigma,1} \leq 3 \sum_{l=1}^{q} \| \Pi_{\mathcal{C}_{\tilde{X}^l}}(X^l - \tilde{X}^l)\|_{\sigma,1} , \tag{2.27}$$

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,1} \leq 4\sqrt{2\,\mathrm{rk}(\tilde{\mathcal{X}})} \sum_{l=1}^{q} \|(X^l - \tilde{X}^l)\|_{\sigma,2} , \tag{2.28}$$

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,1} \leq 4\sqrt{2m_1 m_2\,\mathrm{rk}(\tilde{\mathcal{X}})/K_\gamma}\, \mathrm{d}_{\mathrm{h}}\left(f(\tilde{\mathcal{X}}), f(\mathcal{X}), \right) , \tag{2.29}$$

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,1} \leq 4\sqrt{2m_1 m_2\,\mathrm{rk}(\tilde{\mathcal{X}})/K_\gamma}\sqrt{\mathrm{KL}\left(f(\tilde{\mathcal{X}}), f(\mathcal{X})\right)} . \tag{2.30}$$

Proof. Since $\Phi(\mathcal{X}) \leq \Phi(\tilde{\mathcal{X}})$, we have

$$\Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) - \Phi_{\mathrm{Y}}(\mathcal{X}) \geq \lambda \sum_{l=1}^{q} (\|X^l\|_{\sigma,1} - \|\tilde{X}^l\|_{\sigma,1}).$$

For any $X \in \mathbb{R}^{m_1 \times m_2}$, using $X = \tilde{X} + \Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(X - \tilde{X}) + \Pi_{\mathcal{C}_{\tilde{X}}}(X - \tilde{X})$, Theorem 2.9-(i) and the triangular inequality, we get

$$\|X\|_{\sigma,1} \geq \|\tilde{X}\|_{\sigma,1} + \| \Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(X - \tilde{X})\|_{\sigma,1} - \| \Pi_{\mathcal{C}_{\tilde{X}}}(X - \tilde{X})\|_{\sigma,1} ,$$

which implies

$$\Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) - \Phi_{\mathrm{Y}}(\mathcal{X}) \geq \lambda \sum_{l=1}^{q} \left(\| \Pi_{\mathcal{C}_{\tilde{X}^l}^{\perp}}(X^l - \tilde{X}^l)\|_{\sigma,1} - \| \Pi_{\mathcal{C}_{\tilde{X}^l}}(X - \tilde{X}^l)\|_{\sigma,1} \right) . \tag{2.31}$$

Furthermore by concavity of Φ_{Y} we have

$$\Phi_{\mathrm{Y}}(\tilde{\mathcal{X}}) - \Phi_{\mathrm{Y}}(\mathcal{X}) \leq \sum_{l=1}^{q} \langle \Sigma_Y^l(\tilde{\mathcal{X}}), \tilde{X}^l - X^l \rangle .$$

The duality between $\|\cdot\|_{\sigma,1}$ and $\|\cdot\|_{\sigma,\infty}$ (see for instance (Bhatia, 1997, Corollary IV.2.6)) leads to

$$\begin{aligned}\Phi_Y(\tilde{\mathcal{X}}) - \Phi_Y(\mathcal{X}) &\leq \max_{l\in[q]} \|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty} \sum_{l=1}^{q} \|\tilde{X}^l - X^l\|_{\sigma,1} \ , \\ &\leq \frac{\lambda}{2} \sum_{l=1}^{q} \|\tilde{X}^l - X^l\|_{\sigma,1} \ , \\ &\leq \frac{\lambda}{2} \sum_{l=1}^{q} (\| \Pi_{\mathcal{C}_{\tilde{X}^l}^{\perp}}(X^l - \tilde{X}^l)\|_{\sigma,1} + \| \Pi_{\mathcal{C}_{\tilde{X}^l}}(X^l - \tilde{X}^l)\|_{\sigma,1}) \ , \end{aligned} \tag{2.32}$$

where we used $\lambda > 2\max_{l\in[q]} \|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty}$ in the second line. Then combining (2.31) with (2.32) gives (2.27). Since for any $l \in [q]$, $X^l - \tilde{X}^l = \Pi_{\mathcal{C}_{\tilde{X}^l}^{\perp}}(X^l - \tilde{X}^l) + \Pi_{\mathcal{C}_{\tilde{X}^l}}(X^l - \tilde{X}^l)$, using the triangular inequality and (2.27) yields

$$\sum_{l=1}^{q} \|X^l - \tilde{X}^l\|_{\sigma,1} \leq 4\| \Pi_{\mathcal{C}_{\tilde{X}^l}}(X^l - \tilde{X}^l)\|_{\sigma,1}. \tag{2.33}$$

Combining (2.33) and (2.27) immediately leads to (2.28) and (2.29) is a consequence of (2.28) and the definition of K_γ. The statement (2.30) follows from (2.29) and Theorem 2.10. □

Lemma 2.13. *Under H2 we have*

$$\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \geq \frac{1}{\mu} \mathrm{KL}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \ .$$

where $\mathrm{D}(\cdot,\cdot)$ *is defined in* (2.22).

Proof. Follows from

$$\begin{aligned}\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{\substack{k\in[m_1]\\ l\in[m_2]}} \sum_{j\in[p]} \pi_{k,l} \left[f^j(\bar{\mathcal{X}}_{k,l}) \log\left(\frac{f^j(\bar{\mathcal{X}}_{k,l})}{f^j(\mathcal{X}_{k,l})}\right)\right] \ , \\ &\geq \frac{1}{\mu m_1 m_2} \sum_{\substack{k\in[m_1]\\ l\in[m_2]}} \sum_{j\in[p]} \left[f^j(\bar{\mathcal{X}}_{k,l}) \log\left(\frac{f^j(\bar{\mathcal{X}}_{k,l})}{f^j(\mathcal{X}_{k,l})}\right)\right] \ . \end{aligned}$$

□

Lemma 2.14. *Assume that* $\lambda \geq \bar{\Sigma}$. *Let* $\alpha > 1$, $\beta > 0$ *and* $0 < \eta < 1/2\alpha$. *Then with probability at least*

$$1 - 2(\exp(-n\eta^2 \log(\alpha)\beta^2/(4M_\gamma^2))/(1 - \exp(-n\eta^2 \log(\alpha)\beta^2/(4M_\gamma^2)))$$

we have for all $\mathcal{X} \in \mathcal{C}_\beta(r)$*:*

$$|\, \Phi_Y(\mathcal{X}) - \Phi_Y(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)| \leq \frac{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2} + \epsilon(r,\alpha,\eta) \ ,$$

where

$$\epsilon(r,\alpha,\eta) := \frac{4q^2 L_\gamma^2 r}{1/(2\alpha) - \eta} (\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \ , \tag{2.34}$$

and $\mathcal{C}_\beta(r)$ *is defined in* (2.24).

Proof. The proof is adapted from (Negahban & Wainwright, 2012, Theorem 1) and (Klopp, 2014, Lemma 12). We use a peeling argument combined with a sharp deviation inequality detailed in Theorem 2.15, Consider the events

$$\mathcal{B} := \Bigg\{ \exists \mathcal{X} \in \mathcal{C}_\beta(r) \Bigg| \\ | \Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) | > \frac{\mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right)}{2} + \epsilon(r, \alpha, \eta) \Bigg\} ,$$

and

$$\mathcal{S}_l := \left\{ \mathcal{X} \in \mathcal{C}_\beta(r) | \alpha^{l-1}\beta < \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) < \alpha^l \beta \right\} .$$

Let us also define the set

$$\mathcal{C}_\beta(r, t) = \left\{ \mathcal{X} \in \mathbb{R}^{m_1 \times m_2} |\ \mathcal{X} \in \mathcal{C}_\beta(r),\ \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) \leq t \right\} ,$$

and

$$Z_t := \sup_{\mathcal{X} \in \mathcal{C}_\beta(r,t)} | \Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) | , \tag{2.35}$$

Then for any $\mathcal{X} \in \mathcal{B} \cap \mathcal{S}_l$ we have

$$| \Phi_{\mathrm{Y}}(\mathcal{X}) - \Phi_{\mathrm{Y}}(\bar{\mathcal{X}}) - \mathrm{D}\left(f(\bar{\mathcal{X}}), f(\mathcal{X})\right) | > \frac{1}{2}\alpha^{l-1}\beta + \epsilon(r, \alpha, \eta) ,$$

Moreover by definition of $\mathcal{S}_l$, $\mathcal{X} \in \mathcal{C}_\beta(r, \alpha^l \beta)$. Therefore

$$\mathcal{B} \cap \mathcal{S}_l \subset \mathcal{B}_l := \{ Z_{\alpha^l \beta} > \frac{1}{2\alpha}\alpha^l \beta + \epsilon(r, \alpha, \eta) \} ,$$

If we now apply the union bound and Theorem 2.15 we get

$$\mathbb{P}(\mathcal{B}) \leq \sum_{l=1}^{+\infty} \mathbb{P}(\mathcal{B}_l) \leq \sum_{l=1}^{+\infty} \exp\left(-\frac{n\eta^2(\alpha^l\beta)^2}{8M_\gamma^2} \right) \leq \frac{\exp(-\frac{n\eta^2 \log(\alpha)\beta^2}{4M_\gamma^2})}{1 - \exp(-\frac{n\eta^2 \log(\alpha)\beta^2}{4M_\gamma^2})} ,$$

where we used $x \leq \mathrm{e}^x$ in the second inequality. □

Lemma 2.15. *Assume that $\lambda \geq \bar{\Sigma}$. Let $\alpha > 1$ and $0 < \eta < \frac{1}{2\alpha}$. Then we have*

$$\mathbb{P}\left(Z_t > t/(2\alpha) + \epsilon(r, \alpha, \beta) \right) \leq \exp\left(-n\eta^2 t^2 / (8M_\gamma^2) \right) , \tag{2.36}$$

where Z_t and $\epsilon(r, \alpha, \eta)$ are defined in (2.35) and (2.34), respectively.

Proof. Using Massart's inequality ((Massart, 2000, Theorem 9)) we get for $0 < \eta < 1/(2\alpha)$

$$\mathbb{P}(Z_t > \mathbb{E}[Z_t] + \eta t) \leq \exp\left(-\eta^2 n t^2 / (8M_\gamma^2) \right) . \tag{2.37}$$

By using the standard symmetrization argument, we get

$$\mathbb{E}[Z_t] \leq 2\mathbb{E}\left[\sup_{\mathcal{X} \in \mathcal{C}_\beta(r,t)} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{p} \mathbb{1}_{\{Y_i = j\}} \log\left(\frac{f^j(\mathcal{X}_i)}{f^j(\bar{\mathcal{X}}_i)} \right) \right| \right] ,$$

where $\boldsymbol{\varepsilon} := (\varepsilon_i)_{1\leq i\leq n}$ is a Rademacher sequence which is independent from $(Y_i)_{1\leq i\leq n}$ and $(E_i)_{1\leq i\leq n}$. H5 yields

$$\mathbb{E}[Z_t] \leq \sum_{l=1}^{q} 2\mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\log\left(\frac{g_l^j(X_i^l)}{g_l^j(\bar{X}_i^l)}\right)\right|\right] .$$

Since for any $i\in[n]$, the function

$$\phi_i(x) := \frac{1}{L_\gamma}\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\log\left(\frac{g_l^j(x+\bar{X}_i^l)}{g_l^j(\bar{X}_i^l)}\right)$$

is a contraction satisfying $\phi_i(0)=0$, the contraction principle ((Ledoux & Talagrand, 1991, Theorem 4.12)) and the fact that $(\varepsilon_i)_{i=1}^n$ is independent from $(Y_i)_{i=1}^n$ and $(\omega_i)_{i=1}^n$ yields

$$\mathbb{E}[Z_t] \leq 4L_\gamma\sum_{l=1}^{q}\mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle X^l-\bar{X}^l,E_i\rangle\right|\right] =$$

Denoting $\Sigma_R := n^{-1}\sum_{i=1}^n\varepsilon_iE_i$ and the duality, the previous inequality implies

$$\begin{aligned}\mathbb{E}[Z_t] &\leq 4L_\gamma\sum_{l=1}^{q}\mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)}\left|\langle X^l-\bar{X}^l,\Sigma_R\rangle\right|\right]\\ &\leq 4L_\gamma\sum_{l=1}^{q}\mathbb{E}\left[\sup_{\mathcal{X}\in\mathcal{C}_\beta(r,t)}\|X^l-\bar{X}^l\|_{\sigma,1}\|\Sigma_R\|_{\sigma,\infty}\right] \leq 4qL_\gamma\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt} ,\end{aligned}$$

where we have the definition of $\mathcal{C}_\beta(r,t)$ for the last inequality. Plugging into (2.37) gives

$$\mathbb{P}(Z_t > 4qL_\gamma\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt}+\eta t) \leq \exp\left(-\eta^2nt^2/(8M_\gamma^2)\right) .$$

The proof is concluded by noting that, since for any $a,b\in\mathbb{R}$ and $c>0$, $ab\leq(a^2/c+cb^2)/2$,

$$4qL_\gamma\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt} \leq \frac{1}{1/(2\alpha)-\eta}4q^2L_\gamma^2r\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]^2+(1/(2\alpha)-\eta)t .$$

□

2.5.2 Proof of Theorem 2.7

Proof. By Theorem 2.7 it suffices to control $\|\bar{\Sigma}^l\|_{\sigma,\infty}$ and $\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]$. For any $l\in[q]$, by definition

$$\bar{\Sigma}^l = -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\frac{\partial_l f^j(\bar{\mathcal{X}}_i)}{f^j(\bar{\mathcal{X}}_i)}\right)E_i ,$$

with ∂_l designating the partial derivative against the l-th variable. The sequence of matrices

$$Z_i := \left(\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\frac{\partial_l f^j(\bar{\mathcal{X}}_i)}{f^j(\bar{\mathcal{X}}_i)}\right)E_i = \left(\sum_{j=1}^{p}\mathbb{1}_{\{Y_i=j\}}\frac{(g_l^j)'(\bar{X}_i^l)}{g_l^j(\bar{X}_i^l)}\right)E_i$$

satisfies $\mathbb{E}[Z_i] = 0$ (as any score function) and $\|Z_i\|_{\sigma,\infty} \leq L_\gamma$.
Noticing $e_k(e'_{k'})^\top (e_k(e'_{k'})^\top)^\top = e_k(e'_{k'})^\top$ we also get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i Z_i^\top] = \sum_{k=1}^{m_1}\left(\sum_{k'=1}^{m_2}\pi_{k,k'}\left(\sum_{j=1}^{p} f^j(\bar{\mathcal{X}}_{k,k'})\left(\frac{\partial_l f^j(\bar{\mathcal{X}}_{k,k'})}{f^j(\bar{\mathcal{X}}_{k,k'})}\right)^2\right)\right) e_k(e'_k)^\top \,,$$

which is diagonal. We recall the definition $C_{k'} = \sum_{k=1}^{m_1}\pi_{k,k'}$ and $R_k = \sum_{k'=1}^{m_2}\pi_{k,k'}$ for any $k' \in [m_2]$, $k \in [m_1]$. Since

$$\left(\frac{\partial_l f^j(\bar{\mathcal{X}}_{k,k'})}{f^j(\bar{\mathcal{X}}_{k,k'})}\right)^2 \leq L_\gamma^2 \,,$$

and $(f^j(\bar{\mathcal{X}}_{k,k'}))_{j=1}^p$ is a probability distribution, we obtain

$$\left\|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top\right]\right\|_{\sigma,\infty} \leq L_\gamma^2 \left\|\mathrm{diag}((R_k)_{k=1}^{m_1})\right\|_{\sigma,\infty} \leq L_\gamma^2 \frac{\nu}{m} \,,$$

were we have H3 for the last inequality. Using a similar argument we get $\|\mathbb{E}[\sum_{i=1}^n Z_i^\top Z_i]\|_{\sigma,\infty}/n \leq L_\gamma^2\nu/m$. Therefore, Theorem 2.16 applied with $t = \log(d)$, $U = L_\gamma$ and $\sigma_Z^2 = L_\gamma^2\nu/m$ yields with at least probability $1 - 1/d$,

$$\left\|\Sigma_Y^l(\tilde{X})\right\|_{\sigma,\infty} \leq (1+\sqrt{3})L_\gamma \max\left\{\sqrt{\frac{2\nu\log(d)}{mn}}, \frac{2}{3}\frac{\log(d)}{n}\right\} . \tag{2.38}$$

With the same analysis for $\Sigma_R := \frac{1}{n}\sum_{i=1}^n \varepsilon_i E_i$ and by applying Theorem 2.17 with $U = 1$ and $\sigma_Z^2 = \frac{\nu}{m}$, for $n \geq n^* := m\log(d)/(9\nu)$ it holds:

$$\mathbb{E}\left[\|\Sigma_R\|_{\sigma,\infty}\right] \leq c^* \sqrt{\frac{2e\nu\log(d)}{mn}} \,. \tag{2.39}$$

Assuming $n \geq 2m\log(d)/(9\nu)$, implies $n \geq n^*$ and (2.39) is therefore satisfied. Since it also implies $\sqrt{2\nu\log(d)/(mn)} \geq 2\log(d)/(3n)$, the second term of (2.38) is negligible. Consequently taking $\lambda \geq 2(1+\sqrt{3})L_\gamma\sqrt{2\nu\log(d)/(mn)}$, a union bound argument ensures that $\lambda > 2\max_{l\in[q]}\|\Sigma_Y^l(\tilde{X})\|_{\sigma,\infty}$ with probability at least $1 - q/d$.
By taking λ, β and n as in Theorem 2.7 statement , with probability larger than $1 - (2+q)/d$, Theorem 2.6 result holds when replacing $\mathbb{E}\|\Sigma_R\|_{\sigma,\infty}$ by its upper bound (2.39). Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ yields the result with $\bar{c} = 24832$. □

Proposition 2.16. *Consider a finite sequence of independent random matrices $(Z_i)_{1\leq i\leq n} \in \mathbb{R}^{m_1\times m_2}$ satisfying $\mathbb{E}[Z_i] = 0$ and for some $U > 0$, $\|Z_i\|_{\sigma,\infty} \leq U$ for all $i = 1,\dots,n$. Then for any $t > 0$*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\|_{\sigma,\infty} > t\right) \leq d\exp\left(-\frac{nt^2/2}{\sigma_Z^2 + Ut/3}\right) \,,$$

where $d = m_1 + m_2$ and

$$\sigma_Z^2 := \max\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i Z_i^\top]\right\|_{\sigma,\infty}, \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i^\top Z_i]\right\|_{\sigma,\infty}\right\} .$$

In particular it implies that with at least probability $1 - \mathrm{e}^{-t}$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty} \leq c^* \max \left\{ \sigma_Z \sqrt{\frac{t + \log(d)}{n}}, \frac{U(t + \log(d))}{3n} \right\} ,$$

with $c^* = 1 + \sqrt{3}$.

Proof. The first claim of the proposition is Bernstein's inequality for random matrices (see for example (Tropp, 2012, Theorem 1.6)). Solving the equation (in t) $-\frac{nt^2/2}{\sigma_Z^2 + Ut/3} + \log(d) = -v$ gives with at least probability $1 - \mathrm{e}^{-v}$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty} \leq \frac{1}{n} \left[\frac{U}{3}(v + \log(d)) + \sqrt{\frac{U^2}{9}(v + \log(d))^2 + 2n\sigma_Z^2 (v + \log(d))} \right] ,$$

we conclude the proof by distinguishing the two cases $n\sigma_Z^2 \leq (U^2/9)(v + \log(d))$ or $n\sigma_Z^2 > (U^2/9)(v + \log(d))$. □

Lemma 2.17. *Let* $h \geq 1$. *With the same assumptions as Theorem 2.16, assume* $n \geq (U^2 \log(d))/(9\sigma_Z^2)$ *then the following holds:*

$$\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty}^{h} \right] \leq \left(\frac{2ehc^{*2}\sigma_Z^2 \log(d)}{n} \right)^{h/2} ,$$

with $c^* = 1 + \sqrt{3}$.

Proof. The proof is adapted from (Klopp, 2014, Lemma 6). Define $t^* := (9n\sigma_Z^2)/U^2 - \log(d)$ the value of t for which the two bounds of Theorem 2.16 are equal. Let $\nu_1 := n/(\sigma_Z^2 c^{*2})$ and $\nu_2 := 3n/(Uc^*)$ then, from Theorem 2.16 we have

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty} > t \right) \leq d \exp(-\nu_1 t^2) \text{ for } t \leq t^* ,$$
$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty} > t \right) \leq d \exp(-\nu_2 t) \text{ for } t \geq t^* ,$$

Let $h \geq 1$, then

$$\begin{aligned}
&\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty}^{h} \right] \leq \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty}^{2h\log(d)} \right]^{1/(2\log(d))} , \\
&\leq \left(\int_0^{+\infty} \mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty} > t^{1/(2h\log(d))} \right) dt \right)^{1/(2\log(d))} , \\
&\leq d^{(2h\log(d))^{-1}} \left(\int_0^{+\infty} \exp(-\nu_1 t^{2/(2h\log(d))}) + \exp(-\nu_2 t^{1/(2h\log(d))}) dt \right)^{1/(2\log(d))} , \\
&\leq \sqrt{e} \Big(h\log(d)\nu_1^{-h\log(d)} \Gamma(h\log(d)) + 2h\log(d)\nu_2^{-2h\log(d)} \Gamma(2h\log(d)) \Big)^{1/(2\log(d))} ,
\end{aligned}$$

where we used Jensen's inequality for the first line. Since Gamma-function satisfies for $x \geq 2$, $\Gamma(x) \leq (\frac{x}{2})^{x-1}$ (see (Klopp, 2011, Proposition 12)) we have

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\|_{\sigma,\infty}^{h}\right] \leq$$
$$\sqrt{e}\Big((h\log(d))^{h\log(d)}\nu_1^{-h\log(d)}2^{1-h\log(d)} + 2(h\log(d))^{2h\log(d)}\nu_2^{-2h\log(d)}\Big)^{1/(2\log(d))}.$$

For $n \geq (U^2\log(d))/(9\sigma_Z^2)$ we have $\nu_1\log(d) \leq \nu_2^2$ and therefore we get

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\|_{\sigma,\infty}^{h}\right] \leq \left(\frac{2eh\log(d)}{\nu_1}\right)^{h/2}.$$

□

2.5.3 Proof of Theorem 2.5

Proof. Let h be the following function

$$h(\kappa) = \min\left\{1/2, \sqrt{\alpha\, rMK^{-1}_{(1-\kappa)\gamma}}/(8\gamma\sqrt{n})\right\}. \tag{2.40}$$

Since $0 < h(\kappa) \leq 1/2$ and h is continuous, there exists a fixed point $\kappa^* \in (0, 1/2]$:

$$h(\kappa_*) = \kappa_*. \tag{2.41}$$

For notational convenience, the dependence of κ_* in r, M and n is implicit. We start with a packing set construction, inspired by Davenport et al. (2012). Assume *w.l.o.g.,* that $m_1 \geq m_2$. For $\kappa \leq 1$, define

$$\mathcal{L} = \left\{L = (l_{ij}) \in \mathbb{R}^{m_1\times r} : l_{ij} \in \left\{-\frac{\kappa\gamma}{2}, \frac{\kappa\gamma}{2}\right\}, \forall i \in [m_1],\, \forall j \in [r]\right\},$$

and consider the associated set of block matrices

$$\mathcal{L}' = \left\{L' = (\ L \,\middle|\, \cdots \,\middle|\, L \,\middle|\, O\ ) \in \mathbb{R}^{m_1\times m_2} : L \in \mathcal{L}\right\},$$

where O denotes the $m_1 \times (m_2 - r\lfloor m_2/r\rfloor)$ zero matrix, and $\lfloor x\rfloor$ is the integer part of x.

Remark 2.18. In the case $m_1 < m_2$, we only need to change the construction of the low rank component of the test set. We first build a matrix $\tilde{L} \in \mathbb{R}^{r\times m_2}$ with entries in $\left\{-\frac{\kappa\gamma}{2}, \frac{\kappa\gamma}{2}\right\}$ and then we replicate this matrix to obtain a block matrix L of size $m_1 \times m_2$.

Let $\mathbf{I}_{m_1\times m_2}$ denote the $m_1 \times m_2$ matrix of ones. The Varshamov-Gilbert bound ((Tsybakov, 2009, Lemma 2.9)) guarantees the existence of a subset $\mathcal{L}'' \subset \mathcal{L}'$ with cardinality $\mathrm{Card}(\mathcal{L}'') \geq 2^{(rM)/8} + 1$ containing the matrix $(\kappa\gamma/2)\,\mathbf{I}_{m_1\times m_2}$ and such that, for any two distinct elements X_1 and X_2 of $\mathcal{L}''$,

$$\|X_1 - X_2\|_2^2 \geq \frac{Mr\,\kappa^2\gamma^2}{8}\left\lfloor\frac{m_2}{r}\right\rfloor \geq \frac{m_1 m_2\,\kappa^2\gamma^2}{16}. \tag{2.42}$$

Then, we construct the packing set $\mathcal{A}$ by setting

$$\mathcal{A} = \left\{ L + \frac{(2-\kappa)\gamma}{2} \mathbf{I}_{m_1 \times m_2} \ : \ L \in \mathcal{L}'' \right\}.$$

By construction, any element of $\mathcal{A}$ as well as the difference of any two elements of $\mathcal{A}$ has rank at most r, the entries of any matrix in $\mathcal{A}$ take values in $[0, \gamma]$, and $X^0 = \gamma \mathbf{I}_{m_1 \times m_2}$ belongs to $\mathcal{A}$. Thus, $\mathcal{A} \subset \mathcal{F}(r, \gamma)$. Note that $\mathcal{A}$ has the same size as $\mathcal{L}''$ and it also satisfies the same bound on pairwise distances, i.e. for any two distinct elements X_1 and X_2 of $\mathcal{A}$, (2.42) is satisfied.

For some $X \in \mathcal{A}$, we now estimate the Kullback-Leibler divergence $D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right)$ between probability measures $\mathbb{P}_{X^0}$ and $\mathbb{P}_X$. By independence of the observations $(Y_i, \omega_i)_{i=1}^n$,

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) = n \mathbb{E}_{\omega_1} \left[\sum_{j=1}^{2} f^j(X^0_{\omega_1}) \log\left(\frac{f^j(X^0_{\omega_1})}{f^j(X_{\omega_1})} \right) \right] .$$

Since $X^0_{\omega_1} = \gamma$ and either $X_{\omega_1} = X^0_{\omega_1}$ or $X_{\omega_1} = (1-\kappa)\gamma$, by Theorem 2.19 we get

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \leq \frac{n \left[f^1(\gamma) - f^1\left((1-\kappa)\gamma\right) \right]^2}{f^1\left((1-\kappa)\gamma\right) \left[1 - f^1\left((1-\kappa)\gamma\right) \right]}.$$

From the mean value theorem, for some $\xi \in [(1-\kappa)\gamma, \gamma]$ we have

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \leq \frac{n \{(f^1)'(\xi)\}^2 (\kappa\gamma)^2}{f^1\left((1-\kappa)\gamma\right) \left[1 - f^1\left((1-\kappa)\gamma\right) \right]} .$$

Using H4, the function $(f^1)'$ is decreasing and the latter inequality implies

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \leq 8\, n (\kappa\gamma)^2 g((1-\kappa)\gamma) \ , \tag{2.43}$$

where g is defined in (2.6). From (2.43) and plugging $\kappa = \kappa^*$ defined in eq. (2.41), we get

$$D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \leq \frac{\alpha r M}{8} \leq \alpha \log_2(rM/8) \ ,$$

which implies that

$$\frac{1}{\mathrm{Card}(\mathcal{A}) - 1} \sum_{X \in \mathcal{A}} D\left(\mathbb{P}_{X^0} \| \mathbb{P}_X\right) \ \leq \ \alpha \log\left(\mathrm{Card}(\mathcal{A}) - 1 \right) . \tag{2.44}$$

Using (2.42) and (2.44), (Tsybakov, 2009, Theorem 2.5) implies

$$\inf_{\hat{X}} \sup_{\bar{X} \in \mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}} \left(\frac{\|\hat{X} - \bar{X}\|_2^2}{m_1 m_2} > c \min\left\{ \gamma^2, \frac{Mr}{n\, K_{(1-\kappa^*)\gamma}} \right\} \right) \ \geq \ \delta \tag{2.45}$$

for some universal constants $c > 0$ and $\delta \in (0, 1)$. □

Lemma 2.19. *Let us consider $x, y \in (0,1)$ and*

$$k(x, y) := x \log(x/y) + (1-x) \log((1-x)/1-y) \ .$$

Then the following holds

$$k(x, y) \leq \frac{(x-y)^2}{y(1-y)} \ .$$

Proof. The proof is taken from (Davenport et al., 2012, Lemma 4). Since $k(x, y) = k(1 - x, 1 - y)$, *w.l.o.g.,* we may assume $y > x$. The function $g(t) = k(x, x + t)$ satisfies $g'(t) = t/[(x+t)(1-x-t)]$ and $g''(t) \geq 0$. Therefore the mean value Theorem gives $g(y-x)-g(0) \leq g'(y - x)(y - x)$ which yields the result. □

CHAPTER 3
Low rank matrix completion with exponential family noise

The matrix completion problem consists in reconstructing a matrix from a sample of entries, possibly observed with noise. A popular class of estimator, known as nuclear norm penalized estimators, are based on minimizing the sum of a data fitting term and a nuclear norm penalization. Here, we investigate the case where the noise distribution belongs to the exponential family and is sub-exponential. Our framework alllows for a general sampling scheme. We first consider an estimator defined as the minimizer of the sum of a log-likelihood term and a nuclear norm penalization and prove an upper bound on the Frobenius prediction risk. The rate obtained improves on previous works on matrix completion for exponential family. When the sampling distribution is known, we propose another estimator and prove an oracle inequality *w.r.t.* the Kullback-Leibler prediction risk, which translates immediatly into an upper bound on the Frobenius prediction risk. Finally, we show that all the rates obtained are minimax optimal up to a logarithmic factor.

3.1 Introduction

In the matrix completion problem one aims at recovering a matrix, based on partial and noisy observations of its entries. This problem arises in a wide range of practical situations such as collaborative filtering or quantum tomography (see Srebro & Salakhutdinov (2010) or Gross (2011) for instance). In typical applications, the number of observations is usually much smaller than the total number of entries, so that some structural constraints are needed to recover the whole matrix efficiently.

More precisely, we consider an $m_1 \times m_2$ real matrix $\bar{X}$ and observe n samples of the form $(Y_i, \omega_i)_{i=1}^n$, with $(\omega_i)_{i=1}^n \in ([m_1] \times [m_2])^n$ an $i.i.d.$ sequence of indexes and $(Y_i)_{i=1}^n \in \mathbb{R}^n$ a sequence of observations which is assumed to be $i.i.d.$ conditionally to the entries $(\bar{X}_{\omega_i})_{i=1}^n$. To recover the unknown parameter matrix $\bar{X}$, a popular class of methods, known as penalized nuclear norm estimators, are based on minimizing the sum of a data fitting term and a nuclear norm penalization term. These estimators have been extensively studied over the past decade and strong statistical guarantees can be proved in some particular settings. When the conditional distribution $Y_i|\bar{X}_{\omega_i}$ is additive and sub-exponential it can be shown that the unknown matrix can be recovered efficiently, provided that it is low rank or approximately low rank, see Candès & Plan (2010); Keshavan et al. (2010); Koltchinskii et al. (2011); Negahban & Wainwright (2012); Cai & Zhou (2013a); Klopp (2014). In that case, the prediction error satisfies with high probability

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\frac{(m_1 + m_2)\,\mathrm{rk}(\bar{X})\log(m_1 + m_2)}{n}\right), \tag{3.1}$$

with $\hat{X}$ denoting the estimator, $\|\cdot\|_{\sigma,2}$ the Frobenius norm and $\mathrm{rk}(\cdot)$ the rank of a matrix. It has been proved by Koltchinskii et al. (2011) that this rate is actually minimax optimal up to a logarithmic factor.

Although very common in practice, discrete distributions have received less attention. The analysis of a logistic noise was first addressed by Davenport et al. (2012). It was later considered by Cai & Zhou (2013b), Lafond et al. (2014) and Klopp et al. (2014) who have shown that the prediction error is also of the order of (3.1), for log-likelihood estimators, regularized with nuclear norm. Gunasekar et al. (2014) have investigated the case of distributions belonging to the exponential family, which is rich enough to encompass both continuous and discrete distributions (Gaussian, exponential, Poisson, logistic, etc.). They provide (see their Corollary 1) an upper bound for the prediction error when the noise is sub-Gaussian and the sampling

uniform. However, this bound is of the form

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\alpha^{*2} \frac{(m_1 + m_2)\,\mathrm{rk}(\bar{X})\log(m_1 + m_2)}{n}\right) ,$$

where α^{*2} is of the order $m_1 m_2$ (see Theorem 3.7 below for more details). Therefore, the obtained rate does not match (3.1), which suggests that there may have some room for improvement.

In the present work, we further investigate the case of exponential family distributions and show that under some mild assumptions, the rate (3.1) holds and is minimax optimal up to a logarithmic factor. A matrix completion estimator, defined as the minimizer of the sum of a log-likelihood term and a nuclear norm penalization term, is first considered. Provided that the noise is sub-exponential and the sampling distribution satisfies some assumptions controlling its deviation from the uniform distribution, it is proved that with high probability, the prediction error is upper bounded by the same rate as in the Gaussian setting (3.1). It should be noticed that the sub-exponential assumption is satisfied by all the above mentioned distributions.

When the additional knowledge of the sampling distribution is available, we consider another estimator, which is inspired by the one proposed by Koltchinskii et al. (2011) in the additive sub-exponential noise setting. We adapt their proofs to the exponential family distributions and show that this estimator satisfies an oracle inequality with respect to the Kullback-Leibler prediction risk. The proof techniques involved are also closely related to the dual certificate analysis derived by Zhang & Zhang (2012). With high probability, an upper bound on the prediction error, still of the same order as in (3.1), is derived from the oracle inequality . Finally, it is proved that the previous upper bound order is in fact minimax-optimal up to a logarithmic factor.

The rest of the paper is organized as follows. In Section 3.2.1, the model is specified and some background on exponential family distributions is provided. Then we give an upper bound for log-likelihood matrix completion estimator in Section 3.2.2 and an oracle inequality (also yielding an upper bound) for the estimator with known sampling scheme in Section 3.2.3. Finally, the lower bound is provided in Section 3.2.4. The proofs of the main results are gathered in Section 3.3 and the most technical Lemmas and proofs are deferred to the Appendix.

Notation

Throughout the paper, the following notation will be used. For any integers $n, m_1, m_2 > 0$, $[n] := \{1, \dots, n\}$, $m_1 \vee m_2 := \max(m_1, m_2)$ and $m_1 \wedge m_2 := \min(m_1, m_2)$. We equip the set of $m_1 \times m_2$ matrices with real entries (denoted by $\mathbb{R}^{m_1 \times m_2}$) with the Hilbert-Schmidt inner product $\langle X | X' \rangle := \mathrm{tr}(X^\top X')$. For a given matrix $X \in \mathbb{R}^{m_1 \times m_2}$, we write $\|X\|_\infty := \max_{i,j} |X_{i,j}|$ and for any $s \geq 1$, we denote its Schatten s-norm (see Bhatia (1997)) by

$$\|X\|_{\sigma,s} := \left(\sum_{i=1}^{m_1 \wedge m_2} \sigma_i^s(X)\right)^{1/s} ,$$

with $\sigma_i(X)$ the singular values of X, ordered in decreasing order. We use the convention $\|X\|_{\sigma,\infty} = \sigma_1(X)$. For any vector $z := (z_i)_{i=1}^n$, $\mathrm{diag}(z)$ denotes the $\mathbb{R}^{n \times n}$ diagonal matrix whose diagonal entries are $z_1, \cdots, z_n$. For any convex differentiable function $G : \mathbb{R} \to \mathbb{R}$ and $x, x' \in \mathbb{R}$, the Bregman divergence of G is denoted by

$$d_G(x, x') := G(x) - G(x') - G'(x')(x - x') . \tag{3.2}$$

3.2 Main results

3.2.1 Model Specification

We consider an unknown parameter matrix $\bar{X} \in \mathbb{R}^{m_1 \times m_2}$ that we aim at recovering. Assume that an $i.i.d.$ sequence of indexes $(\omega_i)_{i=1}^n \in ([m_1] \times [m_2])^n$ is sampled and denote by Π its distribution. The observations associated to this sequence are denoted by $(Y_i)_{i=1}^n$ and assumed to follow a natural exponential family distribution, conditionally to the $\bar{X}$ entries, that is:

$$Y_i|\bar{X}_{\omega_i} \sim \mathrm{Exp}_{h,G,(}\,\bar{X}_{\omega_i}) := h(Y_i)\exp\left(\bar{X}_{\omega_i}Y_i - G(\bar{X}_{\omega_i})\right) \ , \tag{3.3}$$

where h and G are the base measure and log partition functions associated to the canonical representation. For ease of notation we often write $\bar{X}_i$ instead of $\bar{X}_{\omega_i}$.

Given two matrices $X^1, X^2 \in \mathbb{R}^{m_1 \times m_2}$, we define the empirical and integrated Bregman divergences as follows

$$D_G^n(X^1, X^2) = \frac{1}{n}\sum_{i=1}^n d_G(X_i^1, X_i^2) \quad \text{and} \quad D_G(X^1, X^2) = \mathbb{E}[D_G^n(X^1, X^2)] \ . \tag{3.4}$$

Note that for exponential family distributions, the Bregman divergence $d_G(\cdot,\cdot)$ corresponds to the Kullback-Leibler divergence. Let $\mathbb{P}_{X^1}$ (*resp.* $\mathbb{P}_{X^2}$) denote the distribution of (Y_1, ω_1) associated to the parameters X^1 (*resp.* X^2); then $D_G^n(X^1, X^2)$ is the Kullback-Leibler divergence between $\mathbb{P}_{X^1}$ and $\mathbb{P}_{X^2}$ conditionally to the sampling, whereas $D_G(X^1, X^2)$ is the usual Kullback-Leibler divergence.

As reminded in introduction, the exponential family encompasses a wide range of distributions, either discrete or continuous. Some information on the most commonly used is recalled below.

Distribution	Parameter x	$G(x)$
Gaussian: $\mathcal{N}(\mu, \sigma^2)$ (σ known)	μ/σ	$\sigma^2 x^2/2$
Binomial: $\mathcal{B}^N(p)$ (N known)	$\log(p/(1-p))$	$N\log(1+\mathrm{e}^x)$
Poisson: $\mathcal{P}(\lambda)$	$\log(\lambda)$	e^x
Exponential: $\mathcal{E}(\lambda)$	$-\lambda$	$-\log(-x)$

TABLE 3.1: Parametrization of some exponential family distributions

Remark 3.1. If G is smooth enough, a simple derivation of the density shows that its successive derivatives can be used to determine the distribution moments. Thus, when G is twice differentiable, $\mathbb{E}[Y_i|\bar{X}_i] = G'(\bar{X}_i)$ and $\mathrm{Var}[Y_i|\bar{X}_i] = G''(\bar{X}_i)$ hold.

3.2.2 General Matrix Completion

In this section, we provide statistical guarantees on the prediction error of a matrix completion estimator, which is defined as the minimizer of the sum of a log-likelihood term and a nuclear norm penalization term. For any $X \in \mathbb{R}^{m_1 \times m_2}$, denote by $\Phi_{\mathrm{Y}}(X)$ the (normalized) conditional negative log-likelihood of the observations:

$$\Phi_{\mathrm{Y}}(X) = -\frac{1}{n}\sum_{i=1}^n \left(\log(h(Y_i)) + X_iY_i - G(X_i)\right) \ . \tag{3.5}$$

For $\gamma > 0$ and $\lambda > 0$, the nuclear norm penalized estimator $\hat{X}$ is defined as follows:

$$\hat{X} = \underset{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_\infty \leq \gamma}{\arg\min} \Phi(X) \ , \quad \text{where} \quad \Phi(X) = \Phi_{\mathrm{Y}}(X) + \lambda \|X\|_{\sigma,1} \ . \tag{3.6}$$

The parameter λ controls the trade off between fitting the data and privileging a low rank solution: for large value of λ, the rank of $\hat{X}$ is expected to be small.

Before giving an upper bound on the prediction risk $\|\hat{X} - \bar{X}\|_{\sigma,2}^2$, the following assumptions on the noise and sampling distributions need to be introduced.

H7. *The function $x \mapsto G(x)$, is twice differentiable and strongly convex on $[-\gamma, \gamma]$, so that there exists constants $\underline{\sigma}, \bar{\sigma} > 0$ satisfying:*

$$\underline{\sigma}^2 \leq G''(x) \leq \bar{\sigma}^2 \ , \tag{3.7}$$

for any $x \in [-\gamma, \gamma]$.

Remark 3.2. Under H7, for any $x, x' \in [-\gamma, \gamma]$, the Bregman divergence satisfies $\underline{\sigma}^2(x-x')^2 \leq 2d_G(x, x') \leq \bar{\sigma}^2(x - x')^2$.

Remark 3.3. If the observations follow a Gaussian distribution, the two convexity constants are equal to the standard deviation *i.e.,* $\bar{\sigma} = \underline{\sigma} = \sigma$ (see Table 3.1).

For the sampling distribution, one needs to ensure that each entry has a sampling probability, which is lower bounded by a strictly positive constant, that is:

H8. *There exists a constant $\mu \geq 1$ such that, for all m_1, m_2,*

$$\min_{k \in [m_1],\, l \in [m_2]} \pi_{k,l} \geq 1/(\mu m_1 m_2) \ , \quad \textit{where } \pi_{k,l} := \mathbb{P}(\omega_1 = (k, l)) \ . \tag{3.8}$$

Denote by $R_k = \sum_{l=1}^{m_2} \pi_{k,l}$ (*resp.* $C_l = \sum_{k=1}^{m_1} \pi_{k,l}$) the probability of sampling a coefficient from row k (*resp.* column l). The following assumption requires that no line nor column should be sampled far more frequently than the others.

H9. *There exists a constant $\nu \geq 1$ such that, for all m_1, m_2,*

$$\max_{k,l}(R_k, C_l) \leq \frac{\nu}{m_1 \wedge m_2} \ .$$

Remark 3.4. In the classical case of a uniform sampling, $\mu = \nu = 1$ holds.

We define the sequence of matrices $(E_i)_{i=1}^n$, whose entries are all zeros except for the coefficient (ω_i) which is equal to one *i.e.,* $E_i := e_{k_i}(e'_{l_i})^\top$ with $(k_i, l_i) = \omega_i$ and $(e_k)_{k=1}^{m_1}$ (*resp.* $(e'_l)_{l=1}^{m_2}$) being the canonical basis of $\mathbb{R}^{m_1}$ (*resp.* $\mathbb{R}^{m_2}$). Furthermore, for $(\varepsilon_i)_{i=1}^n$ a Rademacher sequence independent from $(\omega_i, Y_i)_{i=1}^n$, we also define

$$\Sigma_R := \frac{1}{n} \sum_{i=1}^n \varepsilon_i E_i \ , \tag{3.9}$$

and use the following notation

$$d = m_1 + m_2 \ , \quad M = m_1 \vee m_2, \quad m = m_1 \wedge m_2 \ . \tag{3.10}$$

With these assumptions and notation, we are now ready for stating our main results.

Theorem 3.5. *Assume H7, H8,* $\|\bar{X}\|_\infty \leq \gamma$ *and* $\lambda \geq 2\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$. *Then with probability at least* $1 - 2d^{-1}$ *the following holds:*

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq C\mu^2 \max\left(m_1 m_2 \operatorname{rk}(\bar{X}) \left(\frac{\lambda^2}{\underline{\sigma}^4} + (\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \right), \frac{\gamma^2}{\mu}\sqrt{\frac{\log(d)}{n}} \right) ,$$

with Σ_R *and* d *defined in* (3.9) *and* (3.10) *and* C *a numerical constant.*

Proof. See Section 3.3.1. □

In Theorem 3.5, the term $\mathbb{E}\|\Sigma_R\|_{\sigma,\infty}$ only depends on the sampling distribution and can be upper bounded using assumption H9. On the other hand, the gradient term $\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$ depends both on the sampling and on the observation distributions. In order to control this term with high probability, the noise is assumed to be sub-exponential.

H10. *There exist a constant* $\lambda_\gamma > 0$ *such that for all* $x \in [-\gamma, \gamma]$ *and* $Y \sim \mathrm{Exp}_{h,G,(}\, x)$*:*

$$\mathbb{E}\left[\exp\left(\frac{|Y - G'(x)|}{\lambda_\gamma} \right) \right] \leq \mathrm{e} \quad . \tag{3.11}$$

Then Theorem 3.5, H9 and H10 yield together the following result.

Theorem 3.6. *Assume H7, H8, H9, H10,* $\|\bar{X}\|_\infty \leq \gamma$,

$$n \geq 2\log(d) m \nu^{-1} \max\left(\frac{\lambda_\gamma^2}{\bar{\sigma}^2} \log^2(\lambda_\gamma \sqrt{\frac{m}{\underline{\sigma}^2}}), 1/9 \right) ,$$

and take $\lambda = 2c_\gamma \bar{\sigma}\sqrt{2\nu \log(d)/(mn)}$, *where* c_γ *is a constant which depends only on* λ_γ. *Then with probability at least* $1 - 3d^{-1}$ *the following holds:*

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \bar{C}\mu^2 \max\left[\left(\frac{c_\gamma \bar{\sigma}^2}{\underline{\sigma}^4} + 1 \right) \frac{\nu \operatorname{rk}(\bar{X}) M \log(d)}{n}, \frac{\gamma^2}{\mu}\sqrt{\frac{\log(d)}{n}} \right] ,$$

with $\bar{C}$ *a numerical constant.*

Proof. See Section 3.3.2. □

Remark 3.7. When γ is treated as a constant and n is large, the order of the bound is

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\frac{\operatorname{rk}(\bar{X}) M \log(d)}{n} \right) ,$$

which matches the rate obtained for Gaussian distributions (3.1). Matrix completion for exponential family distributions was considered in the case of uniform sampling (*i.e.,* $\mu = \nu = 1$) and sub-Gaussian noise by Gunasekar et al. (2014). They provide the following upper bound on the estimation error

$$\frac{\|\bar{X} - \hat{X}\|_{\sigma,2}^2}{m_1 m_2} = \mathcal{O}\left(\alpha^{*2} \frac{\operatorname{rk}(\bar{X}) M \log(d)}{n} \right) .$$

with α^* satisfying $\alpha^* \geq \sqrt{m_1 m_2}\|\bar{X}\|_\infty$. Therefore, Theorem 3.6 improves this rate by a factor $m_1 m_2$.

Remark 3.8. In the proof, noncommutative Bernstein inequality for sub-exponential noise is used to control $\|\nabla \Phi_{\mathrm{Y}}(\bar{X})\|_{\sigma,\infty}$. However, when the observations are uniformly bounded (*e.g.,* logistic distribution), a uniform Bernstein inequality can be applied instead, leading in some cases to a sharper bound (see Koltchinskii et al. (2011) and Lafond et al. (2014) for instance).

3.2.3 Matrix Completion with known sampling scheme

When the sampling distribution Π is known, the following estimator can be defined:

$$\check{X} := \underset{X\in\mathbb{R}^{m_1\times m_2},\|X\|_\infty\leq\gamma}{\arg\min} \Phi_{\mathrm{Y}}^{\Pi}(X) + \lambda\|X\|_{\sigma,1} \quad \text{with}\,, \tag{3.12}$$
$$\Phi_{\mathrm{Y}}^{\Pi}(X) := G^{\Pi}(X) - \frac{\sum_{i=1}^n X_i Y_i}{n} \quad \text{and} \quad G^{\Pi}(X) := \mathbb{E}\left[\frac{\sum_{i=1}^n G(X_i)}{n}\right]\,.$$

In the case of sub-exponential additive noise, Koltchinskii et al. (2011) proposed a similar estimator and have shown that it satisfies an oracle inequality *w.r.t.* the Frobenius prediction risk. Note that their estimator coincides with (3.12) for the particular setting of Gaussian noise. The main interest of computing $\check{X}$ instead of $\hat{X}$, when the sampling distribution is known, lies in the fact that a sharp oracle inequality can be derived for $\check{X}$. This powerful tool allows to provide statistical guarantees on the prediction risk, even if the true parameter $\bar{X}$ does not belong to the class of estimators *i.e.,* when $\|\bar{X}\| \leq \gamma$ is not satisfied. In this section, it is proved that $\check{X}$ satisfies an oracle inequality *w.r.t.* the integrated Bregman divergence (see Definition (3.4)), which corresponds to the Kullback-Leibler divergence for exponential family distributions. An upper bound on the Frobenius prediction risk is then easily derived from this inequality.

Theorem 3.9. *Assume H7, H8 and* $\lambda \geq \|\nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})\|_{\sigma,\infty}$. *Then the following inequalities hold:*

$$D_G(\check{X},\bar{X}) \leq \inf_{X\in\mathbb{R}^{m_1\times m_2},\|X\|_\infty\leq\gamma} \left(D_G(X,\bar{X}) + 2\lambda\|X\|_{\sigma,1}\right) \tag{3.13}$$

and

$$D_G(\check{X},\bar{X}) \leq \inf_{X\in\mathbb{R}^{m_1\times m_2},\|X\|_\infty\leq\gamma} \left(D_G(X,\bar{X}) + \left(\frac{1+\sqrt{2}}{2}\right)^2 \frac{\mu}{\underline{\sigma}^2} m_1 m_2 \lambda^2 \operatorname{rk}(X)\right) \tag{3.14}$$

Proof. The proof of Theorem 3.9 is an adaptation (to exponential family distributions) of the proof by Koltchinskii et al. (2011), which uses the first order optimality conditions satisfied by $\check{X}$. Similar arguments are used by Zhang & Zhang (2012) to provide dual certificates for non smooth convex optimization problems. The detailed proof is given in Section 3.6.1. □

When $\|\bar{X}\|_\infty \leq \gamma$, the previous oracle inequalities imply the following upper bound on the prediction risk.

Theorem 3.10. *Assume H7, H8 and* $\lambda \geq \|\nabla \Phi_{\mathrm{Y}}^{\Pi}(\bar{X})\|_{\sigma,\infty}$ *and* $\|\bar{X}\|_\infty \leq \gamma$. *Then the following holds:*

$$\frac{\|\check{X}-\bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \mu^2 \min\left(\frac{\left(1+\sqrt{2}\right)^2}{2}\frac{m_1 m_2}{\underline{\sigma}^4}\lambda^2 \operatorname{rk}(\bar{X}), \frac{4}{\mu\underline{\sigma}^2}\lambda\|\bar{X}\|_{\sigma,1}\right)\,. \tag{3.15}$$

Proof. Applying Theorem 3.9 to $X = \bar{X}$ and using H8 and H7 yields the result. □

As for the previous estimator, the term $\|\nabla \Phi_Y^{\Pi}(\bar{X})\|_{\sigma,\infty}$ is stochastic and depends both on the sampling and observations. Assuming that the sampling distribution is uniform and that the noise is sub-exponential allows to control it with high probability. Before stating the result, let us define

$$L_\gamma := \sup_{x \in [-\gamma,\gamma]} |G'(x)| \,. \tag{3.16}$$

Theorem 3.11. *Assume that the sampling is $i.i.d.$ uniform and $\|\bar{X}\|_\infty \leq \gamma$. Suppose H7, H10, and*

$$n \geq 2\log(d) m \max\left(\frac{\lambda_\gamma^2}{\bar{\sigma}^2} \log^2(\lambda_\gamma \sqrt{\frac{m}{\underline{\sigma}^2}}), 8/9 \right) \,.$$

Take $\lambda = (c_\gamma \bar{\sigma} + c^* L_\gamma)\sqrt{2\log(d)/(mn)}$, where c_γ is a constant which depends only on λ_γ, L_γ is defined in (3.16) and c^* is a numerical constant. Then, with probability at least $1 - 2d^{-1}$ the following holds:

$$\frac{\|\check{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \tilde{C} \left(\frac{c_\gamma \bar{\sigma} + L_\gamma}{\underline{\sigma}^2} \right)^2 \frac{\mathrm{rk}(\bar{X}) M \log(d)}{n} \lambda^2 \,\mathrm{rk}(\bar{X}) \,,$$

with $\tilde{C}$ a numerical constant.

Remark 3.12. For simplicity we have considered here only the case of uniform sampling distributions. However if we assume that the sampling satisfies H8, H9 and that there exists an absolute constant ρ such that $\pi_{k,l} \leq \rho/\sqrt{m_1 m_2}$ for any $m_1, m_2 \in \mathbb{R}$, then it is clear from the proof that the same bound still holds for a general $i.i.d.$ sampling, up to factors depending on μ, ν and ρ.

Remark 3.13. If γ is treated as a constant, the rate obtained for the Frobenius error is the same as in Theorem 3.6. If not, the two rates might differ because the rate of Theorem 3.11 depends on the constant L_γ, which does not appear in Theorem 3.6. Note in addition that Theorem 3.8 also applies to Theorem 3.11.

Proof. The proof is similar to the one of Theorem 3.6, see Section 3.6.2. □

3.2.4 Lower Bound

It can be shown that the upper bounds obtained in Theorems 3.6 and 3.11 are in fact lower bounds (up to a logarithmic factor) when γ is treated as a constant. Before stating the result, let us first introduce the set $\mathcal{F}(r,\gamma)$ of matrices of rank at most r whose entries are bounded by γ:

$$\mathcal{F}(r,\gamma) = \left\{ \bar{X} \in \mathbb{R}^{m_1 \times m_2} : \,\mathrm{rank}(\bar{X}) \leq r,\, \|\bar{X}\|_\infty \leq \gamma \right\} .$$

The infimum over all estimators $\hat{X}$ that are measurable functions of the data $(\omega_i, Y_i)_{i=1}^n$ is denoted by $\inf_{\hat{X}}$.

Theorem 3.14. *There exists two constants $c > 0$ and $\theta > 0$ such that, for all $m_1, m_2 \geq 2$, $1 \leq r \leq m_1 \wedge m_2$, and $\gamma > 0$,*

$$\inf_{\hat{X}} \sup_{\bar{X} \in \mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}} \left(\frac{\|\hat{X} - \bar{X}\|_2^2}{m_1 m_2} > c \min\left\{ \gamma^2, \frac{Mr}{n\,\bar{\sigma}^2} \right\} \right) \geq \theta \,,$$

Remark 3.15. Theorem 3.14 provides a lower bound of order $\mathcal{O}(Mr/(n\,\bar{\sigma}^2)$. The order of the ratio between this lower bound and the upper bounds of Theorem 3.6 is $(c_\gamma(\bar{\sigma}/\underline{\sigma})^4 \log(d) \vee \bar{\sigma}^2)$. If γ is treated as a constant, lower and upper bounds are therefore the same up to a logarithmic factor.

Proof. See Section 3.3.3. □

3.3 Proofs of main results

For $X \in \mathbb{R}^{m_1 \times m_2}$, denote by $\mathcal{S}_1(X) \subset \mathbb{R}^{m_1}$ (*resp.* $\mathcal{S}_2(X) \subset \mathbb{R}^{m_2}$) the linear spans generated by left (*resp.* right) singular vectors of X. Let $P_{\mathcal{S}_1^\perp(X)}$ (*resp.* $P_{\mathcal{S}_2^\perp(X)}$) denotes the orthogonal projections on $\mathcal{S}_1^\perp(X)$ (*resp.* $\mathcal{S}_2^\perp(X)$). We then define the following orthogonal projections on $\mathbb{R}^{m_1 \times m_2}$

$$\Pi_{\mathcal{C}_X^\perp} : \tilde{X} \mapsto P_{\mathcal{S}_1^\perp(X)} \tilde{X} P_{\mathcal{S}_2^\perp(X)} \text{ and } \Pi_{\mathcal{C}X} : \tilde{X} \mapsto \tilde{X} - \Pi_{\mathcal{C}_X^\perp}(\tilde{X}) \; . \tag{3.17}$$

3.3.1 Proof of Theorem 3.5

From Definition (3.6), $f(\hat{X}) \leq f(\bar{X})$ holds, or equivalently

$$D_G^n(\hat{X}, \bar{X}) \leq \lambda(\|\bar{X}\|_{\sigma,1} - \|\hat{X}\|_{\sigma,1}) - \langle \nabla \Phi_{\mathrm{Y}}(\bar{X}), \hat{X} - \bar{X} \rangle \; ,$$

with $D_G^n(\cdot,\cdot)$ defined in (3.4). The first term of the right hand side can be upper bounded using Theorem 3.16-(iii) and the second by duality (between $\|\cdot\|_{\sigma,1}$ and $\|\cdot\|_{\sigma,\infty}$) and the assumption on λ, which yields

$$D_G^n(\hat{X}, \bar{X}) \leq \lambda \left(\|\, \Pi_{\mathcal{C}\,\bar{X}}(\hat{X} - \bar{X})\|_{\sigma,1} + \frac{1}{2}\|\hat{X} - \bar{X}\|_{\sigma,1} \right) \; .$$

Using Theorem 3.16-(ii) to bound the first term and Theorem 3.17-(ii) for the second, leads to

$$D_G^n(\hat{X}, \bar{X}) \leq 3\lambda\sqrt{2\,\mathrm{rk}(\bar{X})}\|\hat{X} - \bar{X}\|_{\sigma,2} \; . \tag{3.18}$$

On the other hand, by strong convexity of G (H7), we get

$$\Delta_{\mathrm{Y}}^2(\hat{X}, \bar{X}) := \frac{1}{n}\sum_{i=1}^n (\hat{X}_i - \bar{X}_i)^2 \leq \frac{2}{\underline{\sigma}^2} D_G^n(\hat{X}, \bar{X}) \; . \tag{3.19}$$

We then define the threshold $\beta := 8\mathrm{e}\gamma^2\sqrt{\log(d)/n}$ and distinguish the two following cases.
Case 1 If $\sum_{kl \in [m_1]\times[m_2]} \pi_{kl}(\hat{X}_{kl} - \bar{X}_{kl})^2 \leq \beta$, then Theorem 3.18 yields

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{m_1 m_2} \leq \mu\beta \; . \tag{3.20}$$

Case 2 If $\sum_{kl \in [m_1]\times[m_2]} \pi_{kl}(\hat{X}_{kl} - \bar{X}_{kl})^2 > \beta$, then Theorem 3.17-(ii) and Theorem 3.18 combined together give

$\hat{X} \in \mathcal{C}(\beta, 32\mu m_1 m_2 \operatorname{rk}(\bar{X}))$, where $\mathcal{C}(\cdot,\cdot)$ is the set defined as

$$\mathcal{C}(\beta, r) := \left\{ X \in \mathbb{R}^{m_1 \times m_2} \middle| \ \|X - \bar{X}\|_{\sigma,1} \leq \sqrt{r\mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]}; \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] > \beta \right\} . \tag{3.21}$$

Hence, from Theorem 3.19 it holds, with probability at least $1 - (d-1)^{-1} \geq 1 - 2d^{-1}$, that

$$\Delta_Y^2(X, \bar{X}) \geq \frac{1}{2}\mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] - 512\mathrm{e}(\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \mu m_1 m_2 \operatorname{rk}(\bar{X}) . \tag{3.22}$$

Combining (3.22) with (3.19), (3.18) and Theorem 3.18 leads to

$$\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}^2}{2\mu m_1 m_2} - 512\mathrm{e}(\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \mu m_1 m_2 \operatorname{rk}(\bar{X}) \leq \frac{6\lambda}{\underline{\sigma}^2}\sqrt{2m_1 m_2 \operatorname{rk}(\bar{X})}\frac{\|\hat{X} - \bar{X}\|_{\sigma,2}}{\sqrt{m_1 m_2}} . \tag{3.23}$$

Using the identity $ab \leq a^2 + b^2/4$ in (3.23) and combining with (3.20) achieves the proof of Theorem 3.5.

Lemma 3.16. *For any pair of matrices X, $\tilde{X} \in \mathbb{R}^{m_1 \times m_2}$ we have*

(i) $\|X + \Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(\tilde{X})\|_{\sigma,1} = \|X\|_{\sigma,1} + \|\Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(\tilde{X})\|_{\sigma,1}$,

(ii) $\|\Pi_{\mathcal{C}_X}(\tilde{X})\|_{\sigma,1} \leq \sqrt{2\operatorname{rk}(X)}\|\tilde{X}\|_{\sigma,2}$,

(iii) $\|X\|_{\sigma,1} - \|\tilde{X}\|_{\sigma,1} \leq \|\Pi_{\mathcal{C}_X}(\tilde{X} - X)\|_{\sigma,1}$.

Lemma 3.17. *Let $X, \tilde{X} \in \mathbb{R}^{m_1 \times m_2}$ satisfying $\|X\|_\infty \leq \gamma$ and $\|\tilde{X}\|_\infty \leq \gamma$. Assume that $\lambda > 2\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$ and $f(X) \leq f(\tilde{X})$. Then*

(i) $\|\Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(X - \tilde{X})\|_{\sigma,1} \leq 3\|\Pi_{\mathcal{C}_{\tilde{X}}}(X - \tilde{X})\|_{\sigma,1}$,

(ii) $\|X - \tilde{X}\|_{\sigma,1} \leq 4\sqrt{2\operatorname{rk}(\tilde{X})}\|(X - \tilde{X})\|_{\sigma,2}$.

Lemma 3.18. *Under H8, for any $X \in \mathbb{R}^{m_1 \times m_2}$ it holds*

$$\sum_{kl \in [m_1] \times [m_2]} \pi_{kl}(X_{kl} - \bar{X}_{kl})^2 \geq \frac{1}{\mu m_1 m_2}\|X - \bar{X}\|_{\sigma,2}^2 .$$

Lemma 3.19. *For $\beta = 8\mathrm{e}\gamma^2\sqrt{\log(d)/n}$, with probability at least $1 - (d-1)^{-1}$, we have for all $X \in \mathcal{C}(\beta, r)$:*

$$\left|\Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]\right| \leq \frac{\mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]}{2} + 16\mathrm{e}(\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 \, r ,$$

with $\mathcal{C}(\beta, r)$ defined in (3.21).

Proof. Theorems 3.16 and 3.17 are proved in Section 3.4. Theorem 3.18 follows directly from H8. See Section 3.5 for the proof of Theorem 3.19. □

3.3.2 Proof of Theorem 3.6

Starting from Theorem 3.5 one only needs to control $\mathbb{E}(\|\Sigma_R\|_{\sigma,\infty})$ and $\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$ to obtain the result.

Control of $\mathbb{E}(\|\Sigma_R\|_{\sigma,\infty})$: One can write $\Sigma_R := n^{-1}\sum_{i=1}^n Z_i$, with $Z_i := \varepsilon_i E_i$ which satisfies $\mathbb{E}[Z_i] = 0$. Recalling the definitions $R_k = \sum_{l=1}^{m_2} \pi_{k,l}$ and $C_l = \sum_{k=1}^{m_1} \pi_{k,l}$ for any $k \in [m_1]$, $l \in [m_2]$, one obtains

$$\left\| \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top \right] \right\|_{\sigma,\infty} \leq \left\| \mathrm{diag}((R_k)_{k=1}^{m_1}) \right\|_{\sigma,\infty} \leq \frac{\nu}{m} , \tag{3.24}$$

where H9 was used for the last inequality. Using a similar argument one also gets $\|\mathbb{E}[\sum_{i=1}^n Z_i^\top Z_i]\|_{\sigma,\infty}/n \leq \nu/m$. Hence applying Theorem 3.20 with $U = 1$ and $\sigma_Z^2 = \nu/m$, for $n \geq m\log(d)/(9\nu)$ yields

$$\mathbb{E}\left[\|\Sigma_R\|_{\sigma,\infty}\right] \leq c^* \sqrt{\frac{2e\nu \log(d)}{mn}} , \tag{3.25}$$

with c^* a numerical constant.

Control of $\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$: Let us define $Z_i' := (Y_i - G'(\bar{X}_i))E_i$, which satisfies $\nabla \Phi_Y(\bar{X}) := n^{-1}\sum_{i=1}^n Z_i'$ and $\mathbb{E}[Z_i'] = 0$ (as any score function) and

$$\sigma_{Z'}^2 := \max\left(\frac{1}{n}\|\mathbb{E}[\sum_{i=1}^n (Z_i')^\top Z_i']\|_{\sigma,\infty} , \frac{1}{n}\|\mathbb{E}[\sum_{i=1}^n Z_i'(Z_i')^\top]\|_{\sigma,\infty} \right) .$$

Using H10, a similar analysis yields $\sigma_{Z'}^2 \leq \bar{\sigma}^2\nu/m$. On the other hand, $\max_{k,l}(R_k, C_l) \geq 1/m$ and $\mathbb{E}[(Y_i - G'(\bar{X}_i))^2] = G''(\bar{X}_i) \geq \underline{\sigma}^2$ gives $\sigma_{Z'}^2 \geq \underline{\sigma}^2/m$. Applying Theorem 3.21 for $t = \log(d)$ gives with probability at least $1 - d^{-1}$

$$\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty} \leq c_\gamma \max\left\{ \bar{\sigma}\sqrt{\nu/m}\sqrt{\frac{2\log(d)}{n}}, \lambda_\gamma \log(\frac{\lambda_\gamma\sqrt{m}}{\underline{\sigma}})\frac{2\log(d)}{n} \right\} , \tag{3.26}$$

with c_γ which depends only on λ_γ. By assumption on n, the left term dominates. Therefore taking λ as in Theorem 3.6 statement yields $\lambda \geq 2\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$ with probability at least $1 - d^{-1}$. A union bound argument combined to Theorem 3.5 achieves Theorem 3.6 proof.

Lemma 3.20. *Consider a finite sequence of independent random matrices $(Z_i)_{1\leq i\leq n} \in \mathbb{R}^{m_1\times m_2}$ satisfying $\mathbb{E}[Z_i] = 0$ and for some $U > 0$, $\|Z_i\|_{\sigma,\infty} \leq U$ for all $i = 1, \ldots, n$ and define*

$$\sigma_Z^2 := \max\left\{ \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i Z_i^\top] \right\|_{\sigma,\infty} , \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i^\top Z_i] \right\|_{\sigma,\infty} \right\} .$$

Then, for any $n \geq (U^2\log(d))/(9\sigma_Z^2)$ the following holds:

$$\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n Z_i \right\|_{\sigma,\infty} \right] \leq c^*\sigma_Z \sqrt{\frac{2e\log(d)}{n}} ,$$

with $c^* = 1 + \sqrt{3}$.

Proof. See Klopp et al. (2014)[Lemma 15]. □

Proposition 3.21. *Consider a finite sequence of independent random matrices* $(Z_i)_{1\leq i\leq n} \in \mathbb{R}^{m_1\times m_2}$ *satisfying* $\mathbb{E}[Z_i] = 0$. *For some* $U > 0$, *assume*

$$\inf\{\lambda > 0 : \mathbb{E}[\exp(\|Z_i\|_{\sigma,\infty}/\lambda)] \leq \mathrm{e}\} \leq U \quad \textit{for} \quad i = 1, \dots, n$$

and define σ_Z *as in Theorem 3.20. Then for any* $t > 0$, *with probability at least* $1 - \mathrm{e}^{-t}$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} Z_i \right\|_{\sigma,\infty} \leq c_U \max\left\{ \sigma_Z \sqrt{\frac{t+\log(d)}{n}}, U\log(\frac{U}{\sigma_Z})\frac{t+\log(d)}{n} \right\},$$

with c_U *a constant which depends only on* U.

Proof. This result is an extension of the sub-exponential noncommutative Bernstein inequality (Koltchinskii, 2013, Theorem 4), to rectangular matrices by dilation, see (Klopp, 2014, Proposition 11) for details. □

3.3.3 Proof of Theorem 3.14

We start with a packing set construction, inspired by Koltchinskii et al. (2011). Assume *w.l.o.g.,* that $m_1 \geq m_2$. Let $\alpha \in (0, 1/8)$ and define $\kappa := \min(1/2, \sqrt{\alpha m_1 r}/(2\gamma\bar{\sigma}^2\sqrt{n})$ and the set of matrices

$$\mathcal{L} = \left\{ L = (l_{ij}) \in \mathbb{R}^{m_1\times r} : l_{ij} \in \{0, \kappa\gamma\}\ , \forall i \in [m_1],\, \forall j \in [r] \right\}.$$

Consider the associated set of block matrices

$$\mathcal{L}' = \left\{ L' = (\ L \,\middle|\, \cdots \,\middle|\, L \,\middle|\, O\ ) \in \mathbb{R}^{m_1\times m_2} : L \in \mathcal{L} \right\},$$

where O denotes the $m_1 \times (m_2 - r\lfloor m_2/r \rfloor)$ zero matrix, and $\lfloor x \rfloor$ is the integer part of x. The Varshamov-Gilbert bound ((Tsybakov, 2009, Lemma 2.9)) guarantees the existence of a subset $\mathcal{A} \subset \mathcal{L}'$ with cardinality $\mathrm{Card}(\mathcal{A}) \geq 2^{(rm_1)/8} + 1$ containing the null matrix X^0 and such that, for any two distinct elements X^1 and X^2 of $\mathcal{A}$,

$$\|X^1 - X^2\|_2^2 \geq \frac{m_1 r\, \kappa^2\gamma^2}{8} \left\lfloor \frac{m_2}{r} \right\rfloor \geq \frac{m_1 m_2\, \kappa^2\gamma^2}{16}. \tag{3.27}$$

By construction, any element of $\mathcal{A}$ as well as the difference of any two elements of $\mathcal{A}$ has rank at most r, the entries of any matrix in $\mathcal{A}$ take values in $[0, \gamma]$ and thus $\mathcal{A} \subset \mathcal{F}(r, \gamma)$. For some $X \in \mathcal{A}$, we now estimate the Kullback-Leibler divergence $D(\mathbb{P}_X \| \mathbb{P}_{X^0})$ between probability measures $\mathbb{P}_{X^0}$ and $\mathbb{P}_X$. By independence of the observations $(Y_i, \omega_i)_{i=1}^n$ and since the distribution of $Y_i|\omega_i$ belongs to the exponential family one obtains

$$D(\mathbb{P}_X \| \mathbb{P}_{X^0}) = n\mathbb{E}_{\omega_1}\left[G'(X_{\omega_1})(X_{\omega_1} - X^0_{\omega_1}) - G(X_{\omega_1}) + G(X^0_{\omega_1}) \right].$$

Since $X^0_{\omega_1} = 0$ and either $X_{\omega_1} = 0$ or $X_{\omega_1} = \kappa\gamma$, by strong convexity and by definition of κ one gets

$$D(\mathbb{P}_X \| \mathbb{P}_{X^0}) \leq n\frac{\bar{\sigma}^2}{2}\kappa^2\gamma^2 \leq \frac{\alpha r m_1}{8} \leq \alpha \log_2(\mathrm{Card}(\mathcal{A}) - 1),$$

which implies

$$\frac{1}{\mathrm{Card}(\mathcal{A}) - 1} \sum_{X\in\mathcal{A}} D(\mathbb{P}_{X^0} \| \mathbb{P}_X) \leq \alpha \log\left(\mathrm{Card}(\mathcal{A}) - 1\right). \tag{3.28}$$

Using (3.27), (3.28) and (Tsybakov, 2009, Theorem 2.5) together gives

$$\inf_{\hat{X}} \sup_{\bar{X}\in\mathcal{F}(r,\gamma)} \mathbb{P}_{\bar{X}}\left(\frac{\|\hat{X}-\bar{X}\|_2^2}{m_1 m_2} > \tilde{c}\min\left\{\gamma^2, \frac{\alpha M r}{n\,\bar{\sigma}^2}\right\}\right) \geq \delta(\alpha, M)\ ,$$

where

$$\delta(\alpha, M) = \frac{1}{1+2^{-rM/16}}\left(1-2\alpha-\frac{1}{2}\sqrt{\frac{\alpha}{rM\log(2)}}\right)\ , \tag{3.29}$$

and $\tilde{c}$ is a numerical constant. Since we are free to choose α as small as possible, this achieves the proof.

3.4 Proof of Theorem 3.16 and Theorem 3.17

Theorem 3.16

Proof. If $A, B \in \mathbb{R}^{m_1\times m_2}$ are two matrices satisfying $\mathcal{S}_i(A) \perp \mathcal{S}_i(B)$, $i = 1, 2$, (see Definition (3.17)) then $\|A+B\|_{\sigma,1} = \|A\|_{\sigma,1} + \|B\|_{\sigma,1}$. Applying this identity with $A = X$ and $B = \Pi_{\mathcal{C}_X^\perp}(\tilde{X})$, we obtain

$$\|X + \Pi_{\mathcal{C}_X^\perp}(\tilde{X})\|_{\sigma,1} = \|X\|_{\sigma,1} + \|\,\Pi_{\mathcal{C}_X^\perp}(\tilde{X})\|_{\sigma,1}\ ,$$

showing (i).

From the definition of $\Pi_{\mathcal{C}_X}(\cdot)$, $\Pi_{\mathcal{C}_X}(\tilde{X}) = P_{\mathcal{S}_1(X)}\tilde{X}P_{\mathcal{S}_2^\perp(X)} + \tilde{X}P_{\mathcal{S}_2(X)}$ holds and therefore $\mathrm{rk}(\Pi_{\mathcal{C}_X}(\tilde{X})) \leq 2\,\mathrm{rk}(X)$. On the other hand, the Cauchy-Schwarz inequality implies that for any matrix A, $\|A\|_{\sigma,1} \leq \sqrt{\mathrm{rk}(A)}\|C\|_{\sigma,2}$. Consequently (ii) follows from

$$\|\,\Pi_{\mathcal{C}_X}(\tilde{X})\|_{\sigma,1} \leq \sqrt{2\,\mathrm{rk}(X)}\|\,\Pi_{\mathcal{C}_X}(\tilde{X})\|_{\sigma,2} \leq \sqrt{2\,\mathrm{rk}(X)}\|\tilde{X}\|_{\sigma,2}\ .$$

Finally, since $\tilde{X} = X + \Pi_{\mathcal{C}_X^\perp}(\tilde{X}-X) + \Pi_{\mathcal{C}_X}(\tilde{X}-X)$ we have

$$\begin{aligned}\|\tilde{X}\|_{\sigma,1} &\geq \|X + \Pi_{\mathcal{C}_X^\perp}(\tilde{X}-X)\|_{\sigma,1} - \|\,\Pi_{\mathcal{C}_X}(\tilde{X}-X)\|_{\sigma,1}\ ,\\ &= \|X\|_{\sigma,1} + \|\,\Pi_{\mathcal{C}_X^\perp}(\tilde{X}-X)\|_{\sigma,1} - \|\,\Pi_{\mathcal{C}_X}(\tilde{X}-X)\|_{\sigma,1}\ ,\end{aligned}$$

leading to (iii). □

Theorem 3.17

Proof. Since $f(X) \leq f(\tilde{X})$, we have

$$\Phi_{\mathrm{Y}}(\tilde{X}) - \Phi_{\mathrm{Y}}(X) \geq \lambda(\|X\|_{\sigma,1} - \|\tilde{X}\|_{\sigma,1}).$$

For any $X \in \mathbb{R}^{m_1\times m_2}$, using $X = \tilde{X} + \Pi_{\mathcal{C}_{\tilde{X}}^\perp}(X-\tilde{X}) + \Pi_{\mathcal{C}_{\tilde{X}}}(X-\tilde{X})$, Theorem 3.16-(i) and the triangular inequality, we get

$$\|X\|_{\sigma,1} \geq \|\tilde{X}\|_{\sigma,1} + \|\,\Pi_{\mathcal{C}_{\tilde{X}}^\perp}(X-\tilde{X})\|_{\sigma,1} - \|\,\Pi_{\mathcal{C}_{\tilde{X}}}(X-\tilde{X})\|_{\sigma,1}\ ,$$

which implies

$$\Phi_{\mathrm{Y}}(\tilde{X}) - \Phi_{\mathrm{Y}}(X) \geq \lambda \left(\| \Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(X - \tilde{X})\|_{\sigma,1} - \| \Pi_{\mathcal{C}_{\tilde{X}}}(X - \tilde{X})\|_{\sigma,1} \right) . \tag{3.30}$$

Furthermore by convexity of Φ_{Y} we have

$$\Phi_{\mathrm{Y}}(\tilde{X}) - \Phi_{\mathrm{Y}}(X) \leq \langle \nabla \Phi_{\mathrm{Y}}(\tilde{X}), \tilde{X} - X \rangle ,$$

which yields by duality

$$\begin{aligned} \Phi_{\mathrm{Y}}(\tilde{X}) - \Phi_{\mathrm{Y}}(X) &\leq \|\nabla \Phi_{\mathrm{Y}}(\tilde{X})\|_{\sigma,\infty} \|\tilde{X} - X\|_{\sigma,1} \leq \frac{\lambda}{2} \|\tilde{X} - X\|_{\sigma,1} , \\ &\leq \frac{\lambda}{2} (\| \Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(X - \tilde{X})\|_{\sigma,1} + \| \Pi_{\mathcal{C}_{\tilde{X}}}(X - \tilde{X})\|_{\sigma,1}) , \end{aligned} \tag{3.31}$$

where we used $\lambda > \|\nabla \Phi_{\mathrm{Y}}(\tilde{X})\|_{\sigma,\infty}$ in the second line. Then combining (3.30) with (3.31) gives (i). Since $X - \tilde{X} = \Pi_{\mathcal{C}_{\tilde{X}}^{\perp}}(X - \tilde{X}) + \Pi_{\mathcal{C}_{\tilde{X}}}(X - \tilde{X})$, using the triangular inequality and (i) yields

$$\|X - \tilde{X}|_{\sigma,1} \leq 4 \| \Pi_{\mathcal{C}_{\tilde{X}}}(X - \tilde{X})\|_{\sigma,1}. \tag{3.32}$$

Combining (3.32) and Theorem 3.16-(i) leads to (ii). □

3.5 Proof of Theorem 3.19

Proof. The proof is adapted from (Negahban & Wainwright, 2012, Theorem 1) and (Klopp, 2014, Lemma 12). We use a peeling argument combined with a sharp deviation inequality detailed in Theorem 3.22. For any $\alpha > 1$, $\beta > 0$ and $0 < \eta < 1/2\alpha$, define

$$\epsilon(r, \alpha, \eta) := \frac{4}{1/(2\alpha) - \eta} (\mathbb{E}\|\Sigma_R\|_{\sigma,\infty})^2 r , \tag{3.33}$$

and consider the events

$$\mathcal{B} := \left\{ \exists X \in \mathcal{C}(\beta, r) \middle| \, |\Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]| > \frac{\mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]}{2} + \epsilon(r, \alpha, \eta) \right\} ,$$

and

$$\mathcal{R}_l := \left\{ X \in \mathcal{C}(\beta, r) | \alpha^{l-1}\beta < \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] < \alpha^l \beta \right\} .$$

Let us also define the set

$$\mathcal{C}(\beta, r, t) := \left\{ X \in \mathcal{C}(\beta, r) | \ \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right] \leq t \right\} ,$$

and

$$Z_t := \sup_{X \in \mathcal{C}(\beta, r, t)} |\Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]| . \tag{3.34}$$

Then for any $X \in \mathcal{B} \cap \mathcal{R}_l$ we have

$$|\Delta_Y^2(X, \bar{X}) - \mathbb{E}\left[\Delta_Y^2(X, \bar{X})\right]| > \frac{1}{2}\alpha^{l-1}\beta + \epsilon(r, \alpha, \eta) ,$$

Moreover by definition of $\mathcal{R}_l$, $X \in \mathcal{C}_\beta(r, \alpha^l \beta)$. Therefore

$$\mathcal{B} \cap \mathcal{R}_l \subset \mathcal{B}_l := \{Z_{\alpha^l \beta} > \frac{1}{2\alpha}\alpha^l\beta + \epsilon(r,\alpha,\eta)\} \ ,$$

If we now apply a union bound argument combined to Theorem 3.22 we get

$$\mathbb{P}(\mathcal{B}) \leq \sum_{l=1}^{+\infty} \mathbb{P}(\mathcal{B}_l) \leq \sum_{l=1}^{+\infty} \exp\left(-\frac{n\eta^2(\alpha^l\beta)^2}{8\gamma^4}\right) \leq \frac{\exp(-\frac{n\eta^2\log(\alpha)\beta^2}{4\gamma^4})}{1-\exp(-\frac{n\eta^2\log(\alpha)\beta^2}{4\gamma^4})} \ ,$$

where we used $x \leq \mathrm{e}^x$ in the second inequality. Choosing $\alpha = \mathrm{e}$, $\eta = (4\mathrm{e})^{-1}$ and β as stated in the Lemma yields the result. □

Lemma 3.22. *Let $\alpha > 1$ and $0 < \eta < \frac{1}{2\alpha}$. Then we have*

$$\mathbb{P}\left(Z_t > t/(2\alpha) + \epsilon(r,\alpha,\eta)\right) \leq \exp\left(-n\eta^2 t^2/(8\gamma^4)\right) \ , \tag{3.35}$$

where $\epsilon(r,\alpha,\eta)$ and Z_t are defined in (3.33) and (3.34).

Proof. From Massart's inequality ((Massart, 2000, Theorem 9)) we get for $0 < \eta < 1/(2\alpha)$

$$\mathbb{P}(Z_t > \mathbb{E}[Z_t] + \eta t) \leq \exp\left(-\eta^2 n t^2/(8\gamma^4)\right) \ . \tag{3.36}$$

A symmetrization argument gives

$$\mathbb{E}[Z_t] \leq 2\mathbb{E}\left[\sup_{X\in\mathcal{C}(\beta,r,t)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (X_i - \bar{X}_i)^2\right|\right] \ ,$$

where $\boldsymbol{\varepsilon} := (\varepsilon_i)_{1\leq i\leq n}$ is a Rademacher sequence independent from $(Y_i,\omega_i)_{i=1}^n$. The contraction principle ((Ledoux & Talagrand, 1991, Theorem 4.12)) yields

$$\mathbb{E}[Z_t] \leq 4\mathbb{E}\left[\sup_{X\in\mathcal{C}(\beta,r,t)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (X_i - \bar{X}_i)\right|\right] = 4\mathbb{E}\left[\sup_{X\in\mathcal{C}(\beta,r,t)} \left|\langle \Sigma_R, X - \bar{X}\rangle\right|\right] \ ,$$

where Σ_R is defined in (3.9). Applying the duality inequality and then plugging into (3.36) gives

$$\mathbb{P}(Z_t > 4\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt} + \gamma^2\eta t) \leq \exp\left(-\eta^2 n t^2/(8\gamma^4)\right) \ .$$

Since for any $a, b \in \mathbb{R}$ and $c > 0$, $ab \leq (a^2/c + cb^2)/2$, the proof is concluded by noting that,

$$4\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]\sqrt{rt} \leq \frac{1}{1/(2\alpha)-\eta} 4\mathbb{E}[\|\Sigma_R\|_{\sigma,\infty}]^2 r + (1/(2\alpha)-\eta)t \ .$$

□

3.6 Proof of Oracle inequalities and Bounds for Completion with known sampling

3.6.1 Proof of Theorem 3.9

Proof. The proof is an extension (to the exponential family case) of the one proposed in (Koltchinskii et al., 2011, Theorem 1). For ease of notation, let us define $H := \nabla \Phi_Y^{\Pi}(\bar{X})$

and the set $\Gamma := \{X \in \mathbb{R}^{m_1 \times m_2} | \, \|X\|_\infty \leq \gamma\}$. In view of Theorem 3.1, one obtains

$$H = \frac{\sum_{i=1}^n Y_i E_i}{n} - \nabla G^\Pi(\bar{X}) = \frac{\sum_{i=1}^n (Y_i E_i - \mathbb{E}[Y_i E_i])}{n} \, .$$

From the definition of $\check{X}$, for any $X \in \Gamma$,

$$G^\Pi(\check{X}) - \frac{\sum_{i=1}^n \check{X}_i Y_i}{n} \leq G^\Pi(X) - \frac{\sum_{i=1}^n X_i Y_i}{n} + \lambda(\|X\|_{\sigma,1} - \|\check{X}\|_{\sigma,1})$$

or equivalently

$$\begin{aligned} &G^\Pi(\check{X}) - G^\Pi(\bar{X}) - \langle \nabla G^\Pi(\bar{X}), \check{X} - \bar{X} \rangle \\ &\quad \leq G^\Pi(X) - G^\Pi(\bar{X}) - \langle \nabla G^\Pi(\bar{X}), X - \bar{X} \rangle + \langle H, \check{X} - X \rangle + \lambda(\|X\|_{\sigma,1} - \|\check{X}\|_{\sigma,1}) \end{aligned}$$

Applying Theorem 3.16 (ii),(iii) and duality yields

$$D_G(\check{X}, \bar{X}) - D_G(X, \bar{X}) \leq \lambda(\|\check{X} - X\|_{\sigma,1} + \|X\|_{\sigma,1} - \|\check{X}\|_{\sigma,1}) \leq 2\lambda \|X\|_{\sigma,1} \, .$$

where we used the assumption $\lambda \geq \|H\|_{\sigma,\infty}$. This proves (3.13).

For (3.14), by definition

$$\check{X} = \operatorname*{arg\,min}_{X \in \mathbb{R}^{m_1 \times m2}} F(X) := G^\Pi(X) - \frac{\sum_{i=1}^n X_i Y_i}{n} + \lambda \|X\|_{\sigma,1} + \delta_\Gamma(X) \, ,$$

where δ_Γ is the indicatrice function of the bounded closed convex set Γ *i.e.,* $\delta_\Gamma(x) = 0$ if $x \in \Gamma$ and $\delta_\Gamma(x) = +\infty$ otherwise. Since F is convex, $\check{X}$ satisfies $0 \in \partial F(\check{X})$ with ∂F denoting the subdifferential of F. It is easily checked that the subdifferential $\partial \delta_\Gamma(\check{X})$ is the normal cone of Γ at the point $\check{X}$. Hence, $0 \in \partial F(\check{X})$ implies that there exists $\check{V} \in \partial \|\check{X}\|_{\sigma,1}$ such that for any $X \in \Gamma$,

$$\langle \nabla G^\Pi(\check{X}), \check{X} - X \rangle - \left\langle \frac{\sum_{i=1}^n Y_i E_i}{n} \, | \, \check{X} - X \right\rangle + \lambda \langle \check{V}, \check{X} - X \rangle \leq 0 \, ,$$

or equivalently

$$\langle \nabla G^\Pi(\check{X}) - \nabla G^\Pi(\bar{X}), \check{X} - X \rangle + \lambda \langle \check{V}, \check{X} - X \rangle \leq \langle H, \check{X} - X \rangle \, .$$

For any $\tilde{x}, \bar{x}, x \in \mathbb{R}$, from the Bregman divergence definition it holds

$$(G'(\check{x}) - G'(\bar{x}))(\check{x} - x) = d_G(x, \check{x}) + d_G(\check{x}, \bar{x}) - d_G(x, \bar{x}) \, . \tag{3.37}$$

In addition, for any $V \in \partial \|X\|_{\sigma,1}$, the subdifferential monotonicity yields $\langle \check{V} - V, \check{X} - X \rangle \geq 0$. Therefore

$$D_G(X, \check{X}) + D_G(\check{X}, \bar{X}) - D_G(X, \bar{X}) \leq \langle H, \check{X} - X \rangle - \lambda \langle V, \check{X} - X \rangle \, . \tag{3.38}$$

In Watson (1992), it is shown that:

$$\partial \|X\|_{\sigma,1} = \{ \sum_{i=1}^r u_i v_i^\top + \Pi_{\mathcal{C}_X^\perp} W \;\; | W \in \mathbb{R}^{m_1 \times m_2}, \; \|W\|_{\sigma,\infty} \leq 1 \} \, , \tag{3.39}$$

where $r := \mathrm{rk}(X)$, u_i (*resp.* v_i) are the left (*resp.* right) singular vectors of X and $\Pi_{\mathcal{C}_X^\perp}$ is defined in (3.17). Denote by $\mathcal{S}_1$ (*resp.* $\mathcal{S}_2$) the space of the left (*resp.* right) singular vectors of

X. For $W \in \mathbb{R}^{m_1 \times m_2}$,

$$\left\langle \sum_{i=1}^{r} u_i v_i^\top + \Pi_{\mathcal{C}_X}^{\perp} W \,|\, \check{X} - X \right\rangle = \left\langle \sum_{i=1}^{r} u_i v_i^\top \,|\, P_{\mathcal{S}_1}(\check{X} - X) P_{\mathcal{S}_1} \right\rangle + \left\langle W \,|\, \Pi_{\mathcal{C}_X}^{\perp}(\check{X}) \right\rangle ,$$

and W can be chosen such that $\langle W, \Pi_{\mathcal{C}_X}^{\perp}(\check{X})\rangle = \|\Pi_{\mathcal{C}_X}^{\perp}(\check{X})\|_{\sigma,1}$ and $\|W\|_{\sigma,\infty} \leq 1$. Taking $V \in \partial\|X\|_{\sigma,1}$ associated to this choice of W (in the sense of (3.39)) and $\|\sum_{i=1}^{r} u_i v_i^\top\|_{\sigma,\infty} = 1$ yield

$$D_G(X, \check{X}) + D_G(\check{X}, \bar{X}) - D_G(X, \bar{X}) + \lambda \|\Pi_{\mathcal{C}_X}^{\perp}(\check{X})\|_{\sigma,1} \\ \leq \langle H, \check{X} - X\rangle + \|P_{\mathcal{S}_1}(\check{X} - X)P_{\mathcal{S}_1}\|_{\sigma,1} . \quad (3.40)$$

The first right hand side term can be upper bounded as follows

$$\langle H, \check{X} - X\rangle = \langle H, \Pi_{\mathcal{C}X}(\check{X} - X)\rangle + \langle H, \Pi_{\mathcal{C}_X}^{\perp}(\check{X})\rangle \\ \leq \|H\|_{\sigma,\infty}(\sqrt{2\,\mathrm{rk}(X)}\|\check{X} - X\|_{\sigma,2} + \|\Pi_{\mathcal{C}_X}^{\perp}(\check{X})\|_{\sigma,1}) , \quad (3.41)$$

where duality and Theorem 3.16-(ii) are used for the inequality. Since $\mathrm{rk}(P_{\mathcal{S}_1}(\check{X} - X)P_{\mathcal{S}_1}) \leq \mathrm{rk}(X)$, the second term satisfies

$$\|P_{\mathcal{S}_1}(\check{X} - X)P_{\mathcal{S}_1}\|_{\sigma,1} \leq \sqrt{\mathrm{rk}\, X}\|\check{X} - X\|_{\sigma,2} . \quad (3.42)$$

Using $\lambda \geq \|H\|_{\sigma,\infty}$, (3.40), (3.41) and (3.42) gives

$$D_G(X, \check{X}) + D_G(\check{X}, \bar{X}) + (\lambda - \|H\|_{\sigma,\infty})\|\Pi_{\mathcal{C}_X}^{\perp}(\check{X})\|_{\sigma,1} \\ \leq D_G(X, \bar{X}) + \lambda(1 + \sqrt{2})\sqrt{\mathrm{rk}(X)}\|\check{X} - X\|_{\sigma,2} . \quad (3.43)$$

By H7 and H8, $\|\check{X} - X\|_{\sigma,2} \leq \underline{\sigma}^{-1}\sqrt{2m_1m_2\mu D_G(X, \check{X})}$, hence

$$D_G(\check{X}, \bar{X}) + (\lambda - \|H\|_{\sigma,\infty})\|\Pi_{\mathcal{C}_X}^{\perp}(\check{X})\|_{\sigma,1} \\ \leq D_G(X, \bar{X}) + (\frac{1 + \sqrt{2}}{2})^2 \underline{\sigma}^{-2} m_1 m_2 \mu \lambda^2 \,\mathrm{rk}(X) , \quad (3.44)$$

proving (3.14). □

3.6.2 proof of Theorem 3.11

Proof. By the triangle inequality,

$$\|H\|_{\sigma,\infty} \leq \left\| \frac{\sum_{i=1}^{n}(Y_i - G'(X_i)E_i}{n} \right\|_{\sigma,\infty} + \left\| \frac{\sum_{i=1}^{n} G'(X_i)E_i}{n} - \mathbb{E}[G'(X_1)E_1] \right\|_{\sigma,\infty} , \quad (3.45)$$

holds. As seen in the proof of Theorem 3.6 (in Section 3.3.2), the first term of the right hand side satisfies (3.26) with probability at least $1 - d^{-1}$. If we define $Z_i = G'(X_i)E_i - \mathbb{E}[G'(X_1)E_1]$, then $\mathbb{E}[Z_i] = 0$ gives $\|Z_i\|_{\sigma,\infty} \leq 2L_\gamma$, with L_γ defined in (3.16). A similar argument to the one used to derive Equation (3.24) yields

$$\left\| \mathbb{E}\left[Z_i^\top Z_i \right] \right\|_{\sigma,\infty} \leq \|\mathbb{E}[(G'(X_i)E_i)(G'(X_i)E_i)^\top]\|_{\sigma,\infty} \leq L_\gamma^2 \frac{1}{m} ,$$

and the same bound holds for $\mathbb{E}[Z_i Z_i^\top]$. Therefore, the uniform version of the noncommutative Bernstein inequality (Theorem 3.23) ensures that with probability at least $1 - d^{-1}$

$$\left\| \frac{\sum_{i=1}^n G'(X_i)E_i}{n} - \mathbb{E}[G'(X_1)E_1] \right\|_{\sigma,\infty} \leq c^* \max\left(\frac{L_\gamma}{\sqrt{m}} \sqrt{\frac{2\log(d)}{n}}, 4L_\gamma \frac{\log(d)}{3n} \right). \quad (3.46)$$

Combining (3.26), (3.46) with the assumption made on n in Theorem 3.11, achieves the proof. □

Proposition 3.23. *Consider a finite sequence of independent random matrices $(Z_i)_{1\leq i\leq n} \in \mathbb{R}^{m_1\times m_2}$ satisfying $\mathbb{E}[Z_i] = 0$ and for some $U > 0$, $\|Z_i\|_{\sigma,\infty} \leq U$ for all $i = 1, \ldots, n$. Then for any $t > 0$*

$$\mathbb{P}\left(\left\| \frac{1}{n}\sum_{i=1}^n Z_i \right\|_{\sigma,\infty} > t \right) \leq d \exp\left(-\frac{nt^2/2}{\sigma_Z^2 + Ut/3} \right),$$

where $d = m_1 + m_2$ and

$$\sigma_Z^2 := \max\left\{ \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i Z_i^\top] \right\|_{\sigma,\infty}, \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i^\top Z_i] \right\|_{\sigma,\infty} \right\}.$$

In particular it implies that with at least probability $1 - \mathrm{e}^{-t}$

$$\left\| \frac{1}{n}\sum_{i=1}^n Z_i \right\|_{\sigma,\infty} \leq c^* \max\left\{ \sigma_Z \sqrt{\frac{t+\log(d)}{n}}, \frac{U(t+\log(d))}{3n} \right\},$$

with $c^* = 1 + \sqrt{3}$.

Proof. The first claim of the proposition is Bernstein's inequality for random matrices (see for example (Tropp, 2012, Theorem 1.6)). Solving the equation (in t) $-\frac{nt^2/2}{\sigma_Z^2 + Ut/3} + \log(d) = -v$ gives with at least probability $1 - \mathrm{e}^{-v}$

$$\left\| \frac{1}{n}\sum_{i=1}^n Z_i \right\|_{\sigma,\infty} \leq \frac{1}{n}\left[\frac{U}{3}(v+\log(d)) + \sqrt{\frac{U^2}{9}(v+\log(d))^2 + 2n\sigma_Z^2(v+\log(d))} \right],$$

we conclude the proof by distinguishing the two cases $n\sigma_Z^2 \leq (U^2/9)(v+\log(d))$ or $n\sigma_Z^2 > (U^2/9)(v+\log(d))$. □

CHAPTER 4
On the Stochastic Frank-Wolfe Algorithms for Convex and Non-convex Optimization

In this paper, the stochastic variants of the classical Frank-Wolfe algorithm are considered. We consider minimizing the regret with a stochastic cost. Our online algorithms only require *simple* iterative updates and a *non-adaptive* step size rule, in contrast to the *hybrid* schemes commonly considered in the literature. The proofs rely on bounding the duality gaps of the online algorithms. Several novel results are derived. The regret bound and anytime optimality for a strongly convex stochastic cost are shown to be as fast as $\mathcal{O}(\log^3 T/T)$ and $\mathcal{O}(\log^2 T/T)$, respectively, where T is the number of rounds played. Moreover, the online projection-free algorithms are shown to converge even when the loss is *non-convex*, i.e., the algorithms find a stationary point to the stochastic cost as $T \to \infty$. Numerical experiments on realistic data sets are presented to support our theoretical claims.

4.1 Introduction

Recently, Frank-Wolfe (FW) algorithm Frank & Wolfe (1956) has become popular for high dimensional constrained optimization. Compared to the projected gradient (PG) algorithm (see Beck & Teboulle (2009a); Juditsky & Nemirovski (2012a,b); Nemirovski et al. (2009)), the FW algorithm (a.k.a. conditional gradient method) is appealing due to its *projection-free* nature. The costly projection step in PG is replaced by a linear optimization in FW. The latter admits a closed form solution for many problems of interests in machine learning. More formally, we consider the following constrained optimization problem

$$\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} f(\boldsymbol{\theta}) \;\text{ s.t. }\; \boldsymbol{\theta} \in \mathcal{C}\ , \tag{4.1}$$

where $\mathcal{C}$ is a convex set in the n-dimensional Euclidean space.

This work focuses on the stochastic variants of the FW and the FW with *away step* (AW) algorithms. At each iteration t, the proposed stochastic FW/AW algorithms update the current iterate $\boldsymbol{\theta}_t$ as in classical FW/AW and a step size is taken according to a non-adaptive rule. The only modification involved is that we use an *approximation of the gradient* $\hat{\nabla}_t f(\boldsymbol{\theta}_t)$ as a surrogate of the true gradient $\nabla f(\boldsymbol{\theta}_t)$ of the expected loss that we attempt to minimize. We establish fast convergence of the algorithms under various conditions.

Fast convergence for stochastic projection-free algorithms have been studied in Garber & Hazan (2015a,b); Lan & Zhou (2014); Hazan & Luo (2016). However, these works considered a 'hybrid' approach that involves solving a regularized linear optimization during the updates Garber & Hazan (2015b); Lan & Zhou (2014); or combining projected gradient algorithms with FW Hazan & Luo (2016). The drawback of these algorithms lies in the extra complexities (in implementation and computation) added to the classical FW algorithm. In particular, their implementation requires the knowledge of some upper bounds of the smoothness or the strong convexity of the unknown objective function. In many machine learning applications, these constants typically depends on the data and sharp upper bounds are rarely available, which results in a slowed effective learning rate (see Section 4.5).

Our first contribution is to show that simple stochastic projection-free methods can achieve on-the-par convergence guarantees as the sophisticated algorithms mentioned above. We consider the two following sets of assumptions: **(a)** the objective $f(\cdot)$ is strongly convex, the optimal solutions lie in the interior of $\mathcal{C}$ (cf. H12, for stochastic FW); **(b)** $\mathcal{C}$ is a polytope (cf. H13, for stochastic AW). We establish that under **(a)** and **(b)**, stochastic FW/AW algorithms converge as $\mathcal{O}(t^{-2\alpha})$ when the approximation error $\|\hat{\nabla}_t f(\boldsymbol{\theta}_t) - \nabla_t f(\boldsymbol{\theta}_t)\|$ behaves as $\mathcal{O}(t^{-\alpha})$ for $\alpha \in (0, 1)$ (see theorem 4).

We then apply these results to the online setting where the gradient is approximated by an online computed aggregated gradient. In particular, we present a set of new results for online FW/AW algorithms in the full information setting, i.e., complete knowledge about the loss function is retrieved at each round Agarwal et al. (2010) (see section 4.2). Our online FW algorithm is similar to the online projection-free method proposed in Hazan & Kale (2012), while the online AW algorithm is new. For online FW algorithms, Hazan & Kale (2012) has proven a regret of $\mathcal{O}(\sqrt{\log^2 t/t})$ for convex and smooth stochastic costs. We improve the regret bound to $\mathcal{O}(\log^3 t/t)$ under the two different sets of assumptions **(a)** and **(b)**. An improved *anytime* optimality bound of $\mathcal{O}(\log^2 t/t)$ (compared to $\mathcal{O}(\sqrt{\log^2 t/t})$ in Hazan & Kale (2012)) is also proven. We emphasize the fact that to implement online FW and AW, there is no need to know an upper bound the smoothness or the strong convexity of the objective function as in Garber & Hazan (2015a,b); Lan & Zhou (2014); Hazan & Luo (2016). This is of crucial importance in applications because the effective convergence rate is not slowed by a pessimistic choice of bounds nor a time consuming hyper parameters optimization. We compare our results to the state-of-the-art in Table 4.1.

	Settings	Regret bound	*Anytime* bound
Garber and Hazan, 2015 Garber & Hazan (2015b)	Hybrid algo., Lipschitz cvx. loss *	$\mathcal{O}(\sqrt{1/t})$	$\mathcal{O}(\sqrt{\log t/t})$
	Hybrid algo., strong cvx. loss *†	$\mathcal{O}(\log t/t)$	$\mathcal{O}(\log t/t)$
Hazan and Kale, 2012 Hazan & Kale (2012)	Simple algo., Lipschitz cvx. loss	$\mathcal{O}(\sqrt{\log^2 t/t})$	$\mathcal{O}(\sqrt{\log^2 t/t})$
	Simple algo., strong cvx. loss	$\mathcal{O}(\sqrt{\log^2 t/t})$	$\mathcal{O}(\sqrt{\log^2 t/t})$
This work	Simple algo., strong cvx. loss, interior point (online FW)	$\mathcal{O}(\log^3 t/t)$	$\mathcal{O}(\log^2 t/t)$
	Simple algo., strong cvx. loss, polytope const. (online AW)	$\mathcal{O}(\log^3 t/t)$	$\mathcal{O}(\log^2 t/t)$

requires: †upper bound of Lipschitz constant, †lower bound of strong convexity constant

TABLE 4.1: Convergence rate comparison. Note that the regret bound for Garber & Hazan (2015b) is given under an adversarial loss setting, while the bounds for Hazan & Kale (2012) and our work are based on a stochastic cost. Depending on the applications (see Section 4.5 & subsection 4.5.1), our regret and anytime bounds can be improved to $\mathcal{O}(\log^2 t/t)$ and $\mathcal{O}(\log t/t)$, respectively.

Finally, we show that the online FW/AW algorithms converge to a stationary point even when the loss is *non-convex*, at a rate of $\mathcal{O}(1/\sqrt{t})$ (see proposition 9). To the best of our knowledge, this is the first convergence rate result for non-convex online optimization with projection-free methods.

To support our claims, we perform numerical experiments on online matrix completion using realistic dataset. The proposed online schemes outperform a simple projected gradient method in terms of running time. The algorithm also demonstrates excellent performance for robust binary classification.

Notation. For any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, \cdots, n\}$. The inner product on an n-dimensional real Euclidian space $\mathbf{E}$ is denoted by $\langle \cdot, \cdot \rangle$ and the associated Euclidian norm by $\|\cdot\|_2$. The space $\mathbf{E}$ is also equipped with a norm $\|\cdot\|$ and its dual norm $\|\cdot\|_\star$. Diameter of the set $\mathcal{C}$ *w.r.t.* $\|\cdot\|_\star$ (*resp.* $\|\cdot\|_2$) is denoted by ρ (*resp.* $\bar{\rho}$), that is

$$\rho := \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\star \text{ and } \bar{\rho} := \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{C}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \tag{4.2}$$

The ith element in a vector $\boldsymbol{x}$ is denoted by $[\boldsymbol{x}]_i$. For any finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinal. Given a set $\mathcal{C}$, $\mathrm{cone}(\mathcal{C})$ denotes the cone generated by $\mathcal{C}$ *i.e.,* the smalles cone containing $\mathcal{C}$. $\mathcal{K}$ is a face of if there exist a direction r s.t $\mathcal{K} = \mathcal{C} \cap \{y | \langle r, y - x \rangle = 0\}$ and $\mathcal{C} \subset \{y | \langle r, y - x \rangle \leq 0\}$ for some $x \in \mathcal{K}$. We denote the set of faces of $\mathcal{C}$ by $\mathrm{faces}(\mathcal{C})$. Moreover, a differentiable function g is said to be L-smooth if, for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbf{E}$,

$$g(\tilde{\boldsymbol{\theta}}) \leq g(\boldsymbol{\theta}) + \langle \nabla g(\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \rangle + (L/2)\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \,. \tag{4.3}$$

We also say g is μ-strongly convex if for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbf{E}$ we get

$$g(\tilde{\boldsymbol{\theta}}) \geq g(\boldsymbol{\theta}) + \langle \nabla g(\boldsymbol{\theta}), \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \rangle + (\mu/2)\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \,. \tag{4.4}$$

Lastly, g is said to be G-Lipschitz if for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbf{E}$,

$$|g(\boldsymbol{\theta}) - g(\tilde{\boldsymbol{\theta}})| \leq G\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_* \,. \tag{4.5}$$

Literature Review

. In addition to the references already mentioned, this work is related to the study of stochastic optimization, e.g., Ghosh & Lam (2015); Nemirovski et al. (2009). Ghosh & Lam (2015) describes a FW algorithm using stochastic approximation and proves that the optimality gap, *i.e.,* converges to zero almost surely; Nemirovski et al. (2009) analyses the stochastic projected gradient method and proves that the convergence rate is $\mathcal{O}(\log t/t)$ under strong convexity and that the optimal solution lies in the interior of $\mathcal{C}$. This is similar to assumption H12 in this paper.

Lastly, most recent works on *non-convex* optimization are based on the stochastic projected gradient descent method Allen-Zhu & Hazan (2016); Ge et al. (2015). Projection-free *non-convex* optimization has only been addressed by a few authors Ghosh & Lam (2015); Ermol'ev & Verchenko (1976). At the time when we finished with the writing, we notice that several authors have published articles pertaining to non-convex FW algorithm, e.g., Lacoste-Julien (2016) achieves the same convergence rate as ours with an adaptive step size, Jiang et al. (2016) considers a different assumption on the smoothness of loss function, Yu et al. (2014) has a slower convergence rate than ours. Nevertheless, none of the above has considered an online optimization setting with time varying objective like ours.

4.2 Problem Setup and Algorithms

We consider the constrained optimization problem given in (4.1) and further assume that the objective function f is differentiable on the constraint set $\mathcal{C}$. Furthermore, we let $\hat{\nabla}_t f(\boldsymbol{\theta}_t)$ be an estimate of $\nabla f(\boldsymbol{\theta}_t)$ at iteration t. The estimate satisfies that for $\epsilon > 0$,

H11. *For some $\alpha \in (0, 1]$, $\sigma \geq 0$ and $K \in \mathbb{Z}_+^\star$. With probability at least $1 - \epsilon$,*

$$\|\hat{\nabla}_t f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\| \leq \sigma(\eta_t^\epsilon / \{K + t - 1\})^\alpha, \ \forall\, t \geq 1 \ , \tag{4.6}$$

where $\eta_t^\epsilon \geq 1$ is an increasing sequence such that the right hand side decreases to 0.

In section 4.4, we shall demonstrate that H11 can be satisfied by a number of application examples pertaining to online learning. In these cases, the typical value for α is 0.5.

Stochastic Frank-Wolfe (S-FW)

The stochastic FW algorithm is a direct generalization of the classical FW algorithm, as summarized in Algorithm 9. It differs from the classical FW algorithm only in the sense that the *gradient approximation* $\hat{\nabla}_t f(\boldsymbol{\theta}_t)$ is used in the linear optimization (Step 3).

Algorithm 9 Stochastic Frank-Wolfe (S-FW).

1: **Initialize:** $\boldsymbol{\theta}_1 \leftarrow 0$
2: **for** $t = 1, \ldots$ **do**
3: Solve the linear optimization:

$$\boldsymbol{a}_t \leftarrow \arg\min_{\boldsymbol{a} \in \mathcal{C}} \left\langle \boldsymbol{a}, \hat{\nabla}_t f(\boldsymbol{\theta}_t) \right\rangle . \tag{4.7}$$

4: Compute $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \gamma_t(\boldsymbol{a}_t - \boldsymbol{\theta}_t)$.
5: **end for**

Stochastic away-step Frank-Wolfe (S-AW)

The stochastic counterpart of the away step algorithm is given in Algorithm 10. By construction, the iterate $\boldsymbol{\theta}_t$ is a convex combination of extreme points of $\mathcal{C}$, referred to as active atoms. We denote by $\mathcal{A}_t$ the set of active atoms and denote by $\alpha_t^{\boldsymbol{a}}$ the positive weight of any active atom $\boldsymbol{a} \in \mathcal{A}_t$ at time t, that is:

$$\boldsymbol{\theta}_t = \textstyle\sum_{\boldsymbol{a} \in \mathcal{A}_t} \alpha_t^{\boldsymbol{a}} \cdot \boldsymbol{a} \quad \text{with} \quad \alpha_t^{\boldsymbol{a}} > 0 \ . \tag{4.8}$$

Algorithm 10 Stochastic away step Frank-Wolfe (S-AW).

1: **Initialize:** $n_0 = 0$, $\boldsymbol{\theta}_1 = 0$, $\mathcal{A}_1 = \emptyset$;
2: **for** $t = 1, \ldots$ **do**
3: Solve the linear optimizations with the aggregated gradient:

$$\boldsymbol{a}_t^{\mathsf{FW}} \leftarrow \arg\min_{\boldsymbol{a}\in\mathcal{C}} \left\langle \boldsymbol{a}, \hat{\nabla}_t f(\boldsymbol{\theta}_t)\right\rangle,\ \boldsymbol{a}_t^{\mathsf{AW}} \leftarrow \arg\max_{\boldsymbol{a}\in\mathcal{A}_t} \left\langle \boldsymbol{a}, \hat{\nabla}_t f(\boldsymbol{\theta}_t)\right\rangle \tag{4.9}$$

4: **if** $\langle \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t, \hat{\nabla}_t f(\boldsymbol{\theta}_t)\rangle \leq \langle \boldsymbol{\theta}_t - \boldsymbol{a}_t^{\mathsf{AW}}, \hat{\nabla}_t f(\boldsymbol{\theta}_t)\rangle$ or $\mathcal{A}_t = \emptyset$ **then**
5: `FW step:` $\boldsymbol{d}_t \leftarrow \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t$, $n_t \leftarrow n_{t-1} + 1$, $\hat{\gamma}_t \leftarrow \gamma_{n_t}$ and $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup \{\boldsymbol{a}_t^{\mathsf{FW}}\}$.
6: **else**
7: $\boldsymbol{d}_t \leftarrow \boldsymbol{\theta}_t - \boldsymbol{a}_t^{\mathsf{AW}}$, $\gamma_{\max} = \alpha_t^{\boldsymbol{a}_t^{\mathsf{AW}}}/(1 - \alpha_t^{\boldsymbol{a}_t^{\mathsf{AW}}})$, cf. (4.8) for definition of $\alpha_t^{\boldsymbol{a}_t^{\mathsf{AW}}}$.
8: **if** $\gamma_{\max} \geq \gamma_{n_t - 1}$ **then**
9: `AW step:` $n_t \leftarrow n_{t-1} + 1$ and $\hat{\gamma}_t \leftarrow \gamma_{n_t}$
10: **else**
11: `Drop step:` $\hat{\gamma}_t \leftarrow \gamma_{\max}$, $n_t \leftarrow n_{t-1}$ and $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \setminus \{\boldsymbol{a}_t^{\mathsf{AW}}\}$
12: **end if**
13: **end if**
14: Compute $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \hat{\gamma}_t \boldsymbol{d}_t$.
15: **end for**

At each iteration, two types of step might be taken. If the condition of line 4 in Algorithm 10 is satisfied, we call the iteration a "FW step", otherwise we call it an "AW step". When a FW step is taken, a new atom $\boldsymbol{a}_t^{\mathsf{FW}}$ is selected (4.9), the current iterate $\boldsymbol{\theta}_t$ is moved towards $\boldsymbol{a}_t^{\mathsf{FW}}$ and the active set is updated accordingly (lines 5 and 14). The selected atom is the (extreme) point of $\mathcal{C}$ which is maximally correlated to the negative aggregated gradient. Note that this step is identical to a usual S-FW iteration. When an "AW step" is taken, a currently active atom $\boldsymbol{a}_t^{\mathsf{AW}}$ is selected (4.9) and the current iterate is moved away from $\boldsymbol{a}_t^{\mathsf{AW}}$ (line 7 and 14). The atom $\boldsymbol{a}_t^{\mathsf{AW}}$ is the active atom which is the most correlated to the current gradient approximation. The intuition is that taking the 'away' step prevents the algorithm from following a 'zig-zag' path when $\boldsymbol{\theta}_t$ is close to the boundary of $\mathcal{C}$ Wolfe (1970). We also have the following interesting observation on n_t:

Lemma 2. *Consider Algorithm 10. We have $n_t \geq t/2$ for all t, where n_t is the number of non-drop steps taken until round t.*

Proof. Except at initialization, the active set is never empty. Indeed, if there is only one active atom left, then its weight is 1. Therfore the condition of line 8 is satisfied and the atom cannot be dropped. Denote by q_t the number of iterations where an atom was dropped up to time t (line 11). As noted above, $n_t + q_t = t$ holds. Since to be dropped, an atom needs to be added to the active set $\mathcal{A}_t$ first, $q_t \leq t/2$ also holds, yielding the result. □

Lastly, we note that the S-AW algorithm is similar to the classical AW algorithm Wolfe (1970). The exception is that a fixed step size rule is adopted due to the stochastic optimization setting considered in this paper.

Remark 1. *As the linear optimization (4.9) enumerates over the active atoms $\mathcal{A}_t$ at round t, the S-AW algorithm is suitable when $\mathcal{C}$ is an atomic (or polytope) set, otherwise $|\mathcal{A}_t|$ may become too large.*

Remark 2 (Linear Optimization.)**.** *The run-time complexity of the S-FW and S-AW algorithms depends on finding efficient solution to the linear optimization step. In many cases, this is*

extremely efficient. For example, when $\mathcal{C}$ is the trace-norm ball, then the linear optimization amounts to finding the top singular vectors of the gradient; see Jaggi (2013) for an overview.

4.3 Convergence Analysis of S-FW and S-AW

The main results for the convergence of S-FW/S-AW are presented in this section together with essential arguments of the proofs.

4.3.1 Convex Optimization

We analyze first Algorithm 9 and Algorithm 10 when the expected loss function f is convex. The following results can be obtained straightforwardly from Jaggi (2013). Define the optimality gap (a.k.a. anytime bound) as

$$h_t := f(\boldsymbol{\theta}_t) - \min_{\boldsymbol{\theta}\in\mathcal{C}} f(\boldsymbol{\theta}) \ .$$

Theorem 3. *Consider the sequence $\{\boldsymbol{\theta}_t\}_{t=1}^{\infty}$ generated by S-FW and S-AW algorithms with the step size rule set as $\gamma_t = 2/(t+1)$. Assume H11 and that $f(\boldsymbol{\theta})$ is convex and L-smooth. Then, the following holds with probability at least $1-\epsilon$:*

$$\begin{aligned} &\textit{(S-FW)}\ \ h_t \leq \frac{2L\bar{\rho}^2 + 2\rho\sigma}{2-\alpha} \cdot \left(\frac{\eta_t^{\epsilon}}{t+1}\right)^{\alpha}, \ \forall\, t \geq 2 \ , \\ &\textit{(S-AW)}\ \ h_t \leq \frac{4L\bar{\rho}^2 + 8\rho\sigma}{2-\alpha} \cdot \left(\frac{\eta_t^{\epsilon}}{t+1}\right)^{\alpha}, \ \forall\, t \geq 2 \ . \end{aligned} \tag{4.10}$$

A detailed proof is provided in subsection 4.6.1 for completeness. We notice that the convergence rate given above can be slow, *i.e.,* when $\alpha = 0.5$, it only gives an $\mathcal{O}(\sqrt{1/t})$ bound on the optimality gap. In practice, there may exist extra structures in the optimization problem (4.1) that can be exploited to give a faster convergence rate. In particular, our following analysis will depend on several geometric condition of the constraint set $\mathcal{C}$.

Denote by $\partial\mathcal{C}$ the boundary set of $\mathcal{C}$. For S-FW (Algorithm 9), we consider

H12. *There is a minimizer $\boldsymbol{\theta}^\star$ of f that lies in the interior of $\mathcal{C}$, i.e., $\delta := \inf_{\boldsymbol{s}\in\partial\mathcal{C}} \|\boldsymbol{s}-\boldsymbol{\theta}^\star\|_2 > 0$.*

While H12 appears to be restrictive, for S-AW (Algorithm 10), we can work with a relaxed condition. To be fully understood, we require the notion of pyramidal width of a polytope which was introduced by Lacoste-Julien & Jaggi (2015). For self consistency, we recall it here and refer to the original paper for a detailed presentation. Let us assume that $\mathcal{C}$ is a polytope and denote the finite set of its extreme points by $\mathcal{A}$. For any point $\boldsymbol{\theta} \in \mathcal{C}$ we denote by $\mathcal{A}_{\boldsymbol{\theta}}$ the set of all proper set of $\boldsymbol{\theta}$, that is:

$$\begin{aligned} \mathcal{A}_{\boldsymbol{\theta}} := \{\mathcal{A}' : \mathcal{A}' \subseteq \mathcal{A} \text{ such that } \boldsymbol{\theta} \in \operatorname{conv}(\mathcal{A}') \text{ and} \\ \boldsymbol{\theta} \text{ is a proper convex combination of } \mathcal{A}'\} \ . \end{aligned} \tag{4.11}$$

For a point $\boldsymbol{\theta} \in \mathcal{C}$, a face $\mathcal{K} \in \operatorname{faces}(\mathcal{C})$ and a feasible direction $\boldsymbol{d} \in \operatorname{cone}(\mathcal{C}-\boldsymbol{\theta}) \setminus \{0\}$ we define the pyramidal directional width of $\mathcal{K}$ with respect to the direction d and base point $\boldsymbol{\theta}$ as:

$$\operatorname{PdirW}(\mathcal{K}, d, \boldsymbol{\theta}) = \min_{\mathcal{A}'\in\mathcal{A}_{\boldsymbol{\theta}}} \frac{1}{\|\boldsymbol{d}\|_2} \Big(\max_{\boldsymbol{y}\in\mathcal{A}'\cup\{\boldsymbol{a}(\mathcal{K},\boldsymbol{d})\}} \langle \boldsymbol{d}, \boldsymbol{y}\rangle - \min_{\boldsymbol{y}\in\mathcal{A}'\cup\{\boldsymbol{a}(\mathcal{K},\boldsymbol{d})\}} \langle \boldsymbol{d}, \boldsymbol{y}\rangle \Big) \ ,$$

where and $\boldsymbol{a}(\mathcal{K}, \boldsymbol{d}) := \arg\max_{\boldsymbol{v}\in\mathcal{K}}\langle \boldsymbol{v}, \boldsymbol{d}\rangle$. Finally, we define δ_{AW}, the pyramidal width of $\mathcal{C}$ by taking the minimum of the previous quantity over all the faces and points of $\mathcal{C}$:

$$\delta_{\text{AW}} := \min_{\mathcal{K}\in\text{faces}(\mathcal{C}),\boldsymbol{\theta}\in\mathcal{K},\boldsymbol{d}\in\text{cone}(\mathcal{C}-\boldsymbol{\theta})\setminus\{0\}} \text{PdirW}(\mathcal{K}, d, \boldsymbol{\theta}) \ . \tag{4.12}$$

In (Lacoste-Julien & Jaggi, 2015, Theorem 6) it is proved that the pyramidal width of a polytope is positive.

H13. *$\mathcal{C}$ is a polytope. Hence its pyramidal δ_{AW} is positive and finite.*

As it is shown in the proofs, δ_{AW} plays a similar role in S-AW as δ in S-FW. Together with the assumption that the *stochastic* objective f is strongly convex, the S-FW and S-AW algorithms have the following fast convergence rate:

Theorem 4. *Consider the sequence $\{\boldsymbol{\theta}_t\}_{t=1}^{\infty}$ generated by S-FW (*resp. *S-AW) algorithm. Assume H11, H12 (*resp. *H13) and that $f(\boldsymbol{\theta})$ is L-smooth, μ-strongly convex. Set $\gamma_t = 2/(t+1)$. With probability at least $1-\epsilon$ and for all $t \geq 1$, we have*

$$\begin{aligned} &\textit{(S-FW)} \quad h_t \leq \left(\max\left\{2\left(\frac{3}{2}\right)^{\alpha}, 1+\frac{2\alpha}{2-\alpha}\right\}\cdot\frac{\sigma\rho + L\bar{\rho}^2}{2\delta\sqrt{\mu}}\right)^2 \cdot \left(\frac{\eta_t^{\epsilon}}{t+1}\right)^{2\alpha}, \\ &\textit{(S-AW)} \quad h_t \leq 2\left(\max\left\{\left(\frac{3}{2}\right)^{\alpha}, 1+\frac{2\alpha}{2-\alpha}\right\}\cdot\frac{2\sigma\rho + L\bar{\rho}^2}{\delta_{\text{AW}}\sqrt{\mu}}\right)^2 \cdot \left(\frac{\eta_t^{\epsilon}}{t+1}\right)^{2\alpha}. \end{aligned} \tag{4.13}$$

Proof. See Appendix 4.6.2. □

When $\alpha = 0.5$, Theorem 4 improves the previous known bound of $h_t = \mathcal{O}(\sqrt{\eta_t^{\epsilon}/t})$ in Theorem 3 with the additional conditions of strong convexity and H12 or H13 to $\mathcal{O}(\eta_t^{\epsilon}/t)$. It also matches the information-theoretical lower bound for strongly convex stochastic optimization in Raginsky & Rakhlin (2011) (up to a log factor). Moreover, for S-AW, the strong convexity requirement on f can be relaxed; see the discussions in subsection 4.6.7.

Sketch of the Proof of Theorem 4

To provide some insights, we present the main ideas behind the proof of Theorem 4. To simplify the discussion we only consider S-FW, $K = 1$, $\eta_t^{\epsilon} = 1$ and $\alpha = 0.5$ in H11. The full proof can be found in the appendix. Let us define the FW gap:

$$g_t := \max_{\boldsymbol{b}\in\mathcal{C}}\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{b}\rangle \ . \tag{4.14}$$

Since $f(\cdot)$ is L-smooth and $\mathcal{C}$ has a diameter of $\bar{\rho}$, we have

$$f(\boldsymbol{\theta}_{t+1}) \leq f(\boldsymbol{\theta}_t) + \gamma_t\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle + \gamma_t^2 L\bar{\rho}^2/2 \tag{4.15}$$

If we define $\boldsymbol{\epsilon}_t := \hat{\nabla}_t f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)$, and subtract $f(\boldsymbol{\theta}^*)$ on both sides of the inequality above, applying Cauchy Schwartz yields

$$h_{t+1} \leq h_t - \gamma_t g_t + \gamma_t^2 L\bar{\rho}^2/2 + \gamma_t\rho\|\boldsymbol{\epsilon}_t\| \ . \tag{4.16}$$

Observe that as $h_t, g_t \geq 0$, the FW gap term g_t determines the convergence rate of the sequence h_t to zero.

In fact, when f is convex, one can prove $g_t \geq h_t - \rho\|\boldsymbol{\epsilon}_t\|$. By the assumption H11, with probability at least $1-\epsilon$, we have

$$h_{t+1} \leq h_t - \gamma_t h_t + \gamma_t^2 L\bar{\rho}^2/2 + 2\gamma_t\rho\sigma/\sqrt{t} = (1-\gamma_t)h_t + \mathcal{O}(t^{-1.5}) \ . \tag{4.17}$$

Setting $\gamma_t = 1/t$ and a simple induction on the above inequality proves $h_t = \mathcal{O}(1/\sqrt{t})$.

An important consequence of H12 is that the latter leads to a tighter lower bound on g_t. As we present in Lemma 13 in subsection 4.6.2, under H12 and when f is μ-strongly convex, we can lower bound g_t as

$$g_t \geq \max\{0, \delta\sqrt{\mu h_t} - \rho\|\boldsymbol{\epsilon}_t\|\}. \tag{4.18}$$

Note that h_t converges to zero and the above lower bound on g_t eventually will become tighter than the previous one, i.e., $g_t \geq \delta\sqrt{\mu h_t} - \rho\|\boldsymbol{\epsilon}_t\| \geq h_t - \rho\|\boldsymbol{\epsilon}_t\|$. This leads to the accelerated convergence of h_t. More formally, plugging the lower bound into (4.16) gives

$$h_{t+1} \leq h_t - \gamma_t\delta\sqrt{\mu h_t} + \gamma_t^2 L\bar{\rho}^2/2 + 2\gamma_t\rho\sigma/\sqrt{t} \ . \tag{4.19}$$

Again, setting $\gamma_t = 1/t$ and a carefully executed induction argument shows $h_t = \mathcal{O}(1/t)$. The same line of arguments is also used to prove the convergence rate of S-AW, where H13 will be required (instead of H12) to provide a similarly tight lower bound to the FW gap.

4.3.2 Non-convex Optimization

When the objective function f is non-convex, the S-FW/S-AW algorithms can still converge to a stationary point of (4.1). Here, we shall work with a slightly more general setting that is useful for analyzing the online variants of FW or AW algorithms. In particular, we allow the objective function $f(\boldsymbol{\theta})$ to be time-varying and be denoted by the sequence $\{f_t(\boldsymbol{\theta})\}_{t=1}^{\infty}$. The stochastic gradient $\hat{\nabla}_t f(\boldsymbol{\theta}_t)$ in S-FW/S-AW satisfies the following assumptions that is similar to H11:

H14. *For some sufficiently large T_0, $\alpha \in (0,1]$ and $\sigma \geq 0$. With probability at least $1-\epsilon$, we have*

$$\|\hat{\nabla}_t f(\boldsymbol{\theta}_t) - \nabla f_t(\boldsymbol{\theta}_t)\| \leq \sigma \cdot t^{-\alpha}, \ \forall\, t \geq T_0/2 \ . \tag{4.20}$$

Moreover, the sequence of objective functions satisfies:

H15. *The sequence $\{f_t(\boldsymbol{\theta})\}_{t=1}^{\infty}$ satisfy $|f_t(\boldsymbol{\theta}) - f_{t-1}(\boldsymbol{\theta})| \leq C_b \cdot t^{-\beta}$ for all $\boldsymbol{\theta} \in \mathcal{C}$.*

We remark that under H15 and if $\beta \leq 1$, the sequence of functions $\{f_t(\boldsymbol{\theta})\}_{t=1}^{\infty}$ does not necessarily converge. The following theorem shows that S-FW/S-AW algorithms converge to a stationary point of (4.1).

Theorem 5. *Choose the step size as $\gamma_t = t^{-\eta}$ for some $\eta \in [0.5, 1)$. Assume H14, H15 and that each of $f_t(\boldsymbol{\theta})$ is L-smooth and bounded by B (possibly non-convex). Then the following hold with probability at least $1-\epsilon$ — (i) for S-FW algorithm and any $T \geq \max\{T_0, 6\}$,*

$$\min_{t\in[T/2+1,T]} \max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle \leq \left(1 - \left(\frac{2}{3}\right)^{1-\eta}\right)^{-1} \left(2B + \left(C_b + 2\rho\sigma + \frac{L\bar{\rho}^2}{2}\right)\cdot \log 2\right) \cdot T^{-\min\{1-\eta, \beta-\eta, \alpha\}} \ , \tag{4.21}$$

and for S-AW algorithm and any $T \geq \max\{T_0, 20\}$,

$$\min_{t\in[T/2+1,T]} \max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle \leq \left(1 - \left(\frac{4}{5}\right)^{1-\eta}\right)^{-1} \left(2B + \left(L\bar{\rho}^2 + 8\rho\sigma + C_b\right)\cdot \log 2\right)\cdot T^{-\min\{1-\eta,\beta-\eta,\alpha\}}\,, \tag{4.22}$$

(ii) In addition, for any $T \geq \max\{T_0, 2\}$, *we have: either (a)* $f_t(\boldsymbol{\theta}_{t+1}) < f_t(\boldsymbol{\theta}_t)$, $\forall t \in [T/2+1, T]$ *or (b-i) for S-FW,*

$$\max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle \leq \frac{2\rho\sigma}{t^\alpha} + \frac{L\bar{\rho}^2}{2t^\eta}, \text{ for some } t \in [T/2+1, T]\,; \tag{4.23}$$

or (b-ii) for S-AW,

$$\max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle \leq \frac{2\rho\sigma}{t^\alpha} + \frac{L\bar{\rho}^2}{2}\hat{\gamma}_t, \text{ for some } t \in [T/2+1, T]\,. \tag{4.24}$$

By further noting that $\hat{\gamma}_t \leq (T/4)^{-\eta}$ *for all* $t \in [T/2+1, T]$, *the above implies that the FW gap bound can be improved to* $\mathcal{O}(1/T^{\min\{\eta,\alpha\}})$ *or the objective value will be monotonically decreasing for the epoch considered;*

(iii) Finally, when $C_b = 0$, i.e., $f_t(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$ *for all* $t \geq 1$, $\eta + \alpha > 1$ *and* $\eta > 0.5$. *Further assume that* $f(\bar{\boldsymbol{\theta}})$ *takes a finite number of values for the stationary points* $\bar{\boldsymbol{\theta}}$, *then the sequence* $\{\boldsymbol{\theta}_t\}_{t\geq 1}$ *has limit points and each limit point* $\underline{\boldsymbol{\theta}}$ *satisfies*

$$\max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f(\underline{\boldsymbol{\theta}}), \underline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle = 0\,. \tag{4.25}$$

Notice that when the FW gap $\max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle$ becomes zero, it implies that $\langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta} - \boldsymbol{\theta}_t\rangle \geq 0$ for all $\boldsymbol{\theta} \in \mathcal{C}$, *i.e.,* $\boldsymbol{\theta}_t$ is a stationary point to the problem $\min_{\boldsymbol{\theta}\in\mathcal{C}} f_t(\boldsymbol{\theta})$. Moreover, when $\beta \geq 1$, $\alpha \geq 0.5$ and $\eta = 0.5$, then (4.21) shows that the S-FW/S-AW algorithms converge at a rate of $\mathcal{O}(1/\sqrt{T})$.

4.4 Application: Online Learning

We consider the *full information* online learning setting introduced by Agarwal et al. (2010); Hazan & Kale (2012). Let $f(\boldsymbol{\theta})$ be the expectation of a continuously differentiable empirical loss functions $f(\boldsymbol{\theta}; \omega_s)$, where ω_s is drawn i.i.d. from a fixed distribution $\mathcal{D}$, *i.e.,* $f(\boldsymbol{\theta}) := \mathbb{E}_{\omega\sim\mathcal{D}}[f(\boldsymbol{\theta}; \omega)]$. Our goal is to minimize f over a bounded convex constraint set $\mathcal{C}$. The expected regret for a sequence of actions $\{\boldsymbol{\theta}_t\}_{t=1}^T$ is

$$\mathcal{R}_t := t^{-1}\textstyle\sum_{s=1}^t f(\boldsymbol{\theta}_s) - \min_{\boldsymbol{\theta}\in\mathcal{C}} f(\boldsymbol{\theta})\,. \tag{4.26}$$

Let $F_t(\boldsymbol{\theta}) = t^{-1}\sum_{s=1}^t f(\boldsymbol{\theta}; \omega_s)$ the aggregated loss. Our Online FW and AW algorithms (O-FW and O-AW) are obtained by approximating the gradient using the aggregated gradient in Algorithm 9 and Algorithm 10, *i.e.,* we apply S-FW and S-AW with

$$\hat{\nabla}_t f(\boldsymbol{\theta}_t) := \nabla F_t(\boldsymbol{\theta}_t)\,. \tag{4.27}$$

In particular, we observe that $\nabla F_t(\boldsymbol{\theta}_t)$ satisfies the following:

Proposition 6. *Assume that $f(\boldsymbol{\theta};\omega)$ is L-smooth for all ω from $\mathcal{D}$ and each of $\nabla f(\boldsymbol{\theta};\omega_t)$ is sub-Gaussian with parameter σ_D. With probability at least $1-\epsilon$,*

$$\|\nabla F_t(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|_\infty = \mathcal{O}\left(\max(L\bar{\rho},\ \sigma_D)\sqrt{\frac{n\log(t)\log(nt/\epsilon)}{t}}\right),\ \forall\, t \geq 1\,. \tag{4.28}$$

The proof can be found in Section 4.6.4. This shows that $\nabla F_t(\boldsymbol{\theta}_t)$ is an *inexact* gradient of the stochastic objective $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_t$.

Both O-FW/O-AW algorithms require the aggregate gradient $\nabla F_t(\boldsymbol{\theta}_t)$ to be computed at each round, and the complexity involved grows with the round number. In cases when the loss f_t is the negated log-likelihood of an exponential family distribution, the gradient aggregation can be replaced by an efficient 'on-the-fly' update, whose complexity is a dimension-dependent **constant over the iterations**. As demonstrated in Section 4.5 and subsection 4.5.1, this set-up covers many problems of interest, among others the online matrix completion and online LASSO. In the following, we show that the results from the previous section can be applied to derive the regret bounds (or convergence rate) of O-FW and O-AW for convex and non-convex problems.

4.4.1 Convex Loss

When the objective function f is convex, applying Proposition 6 and Theorem 3 shows that for O-FW and O-AW algorithms, we have $h_t = \mathcal{O}(\sqrt{\log^2 t/t})$. This leads to a regret bound of $\mathcal{R}_t = \mathcal{O}(\sqrt{\log^2 t/t})$ which matches the bound in Hazan & Kale (2012). Furthermore, applying Proposition 6 and Theorem 4 shows that the regret bound can be accelerated to $\mathcal{O}(\log^3 t/t)$ under a strongly convex objective and H12 (*resp.* H13) for the O-FW (*resp.* O-AW) algorithm.

Proposition 7. *Assume H12 (or H13), $f(\boldsymbol{\theta})$ is μ-strongly convex, $f(\boldsymbol{\theta};\omega)$ is L-smooth for all ω drawn from $\mathcal{D}$ and each of $\nabla f(\boldsymbol{\theta};\omega_t)$ is sub-Gaussian with parameter σ_D. Set $\gamma_t = 2/(t+1)$. With probability at least $1-\epsilon$ and for all $t \geq 1$, the anytime bounds hold:*

$$\begin{aligned}
&\textit{(O-FW)} \quad h_t \leq \left(\frac{2\sqrt{3/2}(\sigma_{grd}\rho + L\bar{\rho}^2)}{2\delta\sqrt{\mu}}\right)^2 \cdot (\log t \log(nt/\epsilon)) \cdot t^{-1}\,, \\
&\textit{(O-AW)} \quad h_t \leq \left(\frac{(5/3)(2\sigma_{grd}\rho + L\bar{\rho}^2)}{\delta_{\text{AW}}\sqrt{\mu}}\right)^2 \cdot (\log t \log(nt/\epsilon)) \cdot t^{-1}\,,
\end{aligned} \tag{4.29}$$

where $\sigma_{grd} = \mathcal{O}(\max\{L\bar{\rho}, \sigma_D\} \cdot \sqrt{n})$. Consequently, summing up the two sides of (4.29) from $t=1$ to $t=T$ gives the regret bound for both algorithms:

$$\mathcal{R}_T = T^{-1}\textstyle\sum_{t=1}^T h_t = \mathcal{O}(\log^3 T/T),\ \forall\, T \geq 1\,. \tag{4.30}$$

Proof. Our proof is achieved by applying Theorem 4 and Proposition 6 using the appropriate constants. □

We notice that for O-FW, the regret bound of $\mathcal{O}(\sqrt{\log^2 t/t})$ in Hazan & Kale (2012) was proven by using a uniform approximation bound on the *objective function* combined with a $\mathcal{O}(1/\sqrt{t})$ bound for the instantaneous loss $F_t(\boldsymbol{\theta}_t) - \min_{\boldsymbol{\theta}\in\mathcal{C}} F_t(\boldsymbol{\theta})$. This is different from the approach taken in this paper, where we have shown an improved regret bound by controlling

the *gradient error* directly using Proposition 6 and analyzing the O-FW/O-AW algorithms as S-FW/S-AW algorithms studied in the previous section.

4.4.2 Non-convex Loss

We also show convergence of the O-FW/O-AW algorithms for general Lipschitz and smooth (possibly non-convex) objective function $f(\boldsymbol{\theta})$. In this case, note that finding the global minimum $\boldsymbol{\theta}_\star \in \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} f(\boldsymbol{\theta})$ is generally hard as the algorithms may get stuck at a local minimum. Consequently, the regret $\mathcal{R}_t$ may become infinity as $t \to \infty$. As such, our goal is to characterize the convergence (rate) of O-FW/O-AW algorithms to a stationary point of the time varying objective by applying Theorem 5.

Proposition 8. *Consider O-FW and O-AW algorithms. Assume that each of the loss function $f(\boldsymbol{\theta};\omega_t)$ is L-smooth and bounded by B for all $\boldsymbol{\theta} \in \mathcal{C}$. Setting the step size sequence as $\gamma_t = t^{-\eta}$ with $\eta \in [0.5, 1)$. We have that for all $T \geq 20$,*

$$\min_{t\in[T/2+1,T]} \max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla F_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle = \mathcal{O}(1/T^{1-\eta}) \tag{4.31}$$

Proof. The O-FW/O-AW algorithms can be regarded as the S-FW/S-AW algorithms operating on a sequence of objective functions $\{F_t(\boldsymbol{\theta})\}_{t=1}^{\infty}$. Let $\mathrm{cl}(\mathcal{C})$ be the closure of $\mathcal{C}$ and $B < \infty$ be some positive constant. As for each s, the objective function is bounded such that $|f(\boldsymbol{\theta};\omega_s)| \leq B$, we so have $|F_t(\boldsymbol{\theta})| \leq B$. Observe that for all $\boldsymbol{\theta} \in \mathcal{C}$,

$$\begin{aligned} |F_t(\boldsymbol{\theta}) - F_{t-1}(\boldsymbol{\theta})| &= \frac{1}{t} f(\boldsymbol{\theta};\omega_t) + \sum_{s=1}^{t-1} \Big(\frac{1}{t} - \frac{1}{t-1}\Big) f(\boldsymbol{\theta};\omega_s) \\ &= \frac{1}{t}\big(f(\boldsymbol{\theta};\omega_t) - F_{t-1}(\boldsymbol{\theta})\big) \leq \frac{2B}{t},\ \forall\, t \geq 2\ , \end{aligned} \tag{4.32}$$

Clearly, the sequence of function satisfies H15 with $\beta = 1$ and $C_b = 2B$.

For H14, since the O-FW/O-AW algorithms uses the *exact* gradient at round t with respect to the *time varying objective* $F_t(\boldsymbol{\theta}_t)$, assumption H14 is satisfied with $\sigma = 0$, $\alpha = 1$. Finally, by noting that $F_t(\boldsymbol{\theta})$ is also L-smooth, applying (4.21), (4.22) in Theorem 5 yield the desirable result. □

We emphasize that the result above is deterministic and holds with probability one. The drawback is that the stationary point condition is defined with respect to the time-varying objective $F_t(\boldsymbol{\theta})$ and the algorithms may not converge asymptotically. Next, we show that with a slightly more restrictive choice of the step size parameter η, we can have a stronger result:

Proposition 9. *Consider the O-FW/O-AW algorithms, assume that each of $f(\boldsymbol{\theta};\omega_t)$ is bounded by B over $\boldsymbol{\theta} \in \mathcal{C}$ and is L-smooth. Set the step size sequence as $\gamma_t = t^{-\eta}$ with $\eta \in [0.5, 1)$. For any fixed $\epsilon > 0$, $\delta > 0$ and some sufficiently large T_0, it holds with probability at least $1 - \epsilon$ that*

$$\min_{t\in[T/2+1,T]} \max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle = \mathcal{O}\big(1/T^{\min\{0.5-\delta, 1-\eta\}}\big),\ \forall\, T \geq T_0\ . \tag{4.33}$$

Moreover, with probability at least $1 - \epsilon$, if $\eta > 0.5 + \delta$ and $f(\bar{\boldsymbol{\theta}})$ takes a finite number of values for the stationary points $\bar{\boldsymbol{\theta}}$, then the sequence of iterates $\{\boldsymbol{\theta}_t\}_{t=1}^{\infty}$ of the O-FW/O-AW algorithms has limit points and each limit point $\underline{\boldsymbol{\theta}}$ is a stationary point.

Proof. We apply the results from Theorem 5 by considering the O-FW/O-AW algorithms as the S-FW/S-AW algorithms operating on a fixed objective $f(\boldsymbol{\theta})$ with inexact gradient. In particular, under this model, H15 is satisfied with $C_b = 0$ and $\beta = 1$. To satisfy H14, observe that each of $f(\boldsymbol{\theta}; \omega_t)$ is bounded and smooth, this implies that $\nabla f(\boldsymbol{\theta}; \omega_t)$ is also bounded and therefore sub-Gaussian. Since for some sufficiently large T_0, we have $\log t \leq t^{\delta}$ for all $t \geq T_0$, applying Proposition 6 gives

$$\|\nabla F_t(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|_{\infty} = \mathcal{O}\left(\max(L\bar{\rho},\ \sigma_D)\sqrt{n\log(n/\epsilon)} \cdot t^{-0.5+\delta}\right),\ \forall\, t \geq T_0/2\ . \quad (4.34)$$

We see that H14 holds with $\alpha = 0.5 - \delta$. The FW gap bound (4.33) follows from (4.21) and (4.22) (cf. part (a)) in Theorem 5; and the asymptotic convergence follows from the part (c) of Theorem 5 since $\eta + \alpha > 1$. □

4.5 Numerical Experiments

We conduct numerical experiments to demonstrate the practical performance of the online algorithms.

4.5.1 Example: Online LASSO

Consider the setting where we are sequentially given i.i.d. observations $(\boldsymbol{Y}_t, \boldsymbol{A}_t)$ such that $\boldsymbol{Y}_t \in \mathbb{R}^m$ is the response, $\boldsymbol{A}_t \in \mathbb{R}^{m\times n}$ is the random design and $\boldsymbol{Y}_t = \boldsymbol{A}_t\bar{\boldsymbol{\theta}} + \boldsymbol{w}_t$ where the vector $\boldsymbol{w}_t$ is i.i.d., $[\boldsymbol{w}_t]_i$ is independent of $[\boldsymbol{w}_t]_j$ for $i \neq j$ and $[\boldsymbol{w}_t]_i$ is zero-mean and sub-Gaussian with parameter σ_w. We suppose that the unknown parameter $\bar{\boldsymbol{\theta}}$ is sparse. Attempting to learn $\bar{\boldsymbol{\theta}}$, a natural choice for the loss function at round t is the square loss, *i.e.,*

$$f_t(\boldsymbol{\theta}) = (1/2)\|\boldsymbol{Y}_t - \boldsymbol{A}_t\boldsymbol{\theta}\|_2^2 \quad (4.35)$$

and the stochastic cost associated is $f(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}_{\bar{\boldsymbol{\theta}}}[\|\boldsymbol{Y}_t - \boldsymbol{A}_t\boldsymbol{\theta}\|_2^2]$. As $\bar{\boldsymbol{\theta}}$ is sparse, the constraint set is designed to be the ℓ_1 ball, i.e., $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^n\ :\ \|\boldsymbol{\theta}\|_1 \leq r\}$, where $r > 0$ is a regularization constant. Note that $\mathcal{C}$ is a polytope.

The aggregated gradient can be expressed as

$$\nabla F_t(\boldsymbol{\theta}_t) = t^{-1}\Big(\textstyle\sum_{s=1}^t \boldsymbol{A}_s^\top \boldsymbol{A}_s\Big)\boldsymbol{\theta}_t - t^{-1}\Big(\textstyle\sum_{s=1}^t \boldsymbol{A}_s^\top \boldsymbol{Y}_s\Big)\ . \quad (4.36)$$

Similar to the case of online matrix completion, the terms $\sum_{s=1}^t \boldsymbol{A}_s^\top \boldsymbol{A}_s$ and $\sum_{s=1}^t \boldsymbol{A}_s^\top \boldsymbol{Y}_s$ can be computed 'on-the-fly' as running sums. Applying O-FW (9) or O-AW (10) with the above aggregated gradient yields an online LASSO algorithm with a constant complexity (dimension-dependent) per iteration. Notice that as $\mathcal{C}$ is an ℓ_1 ball constraint, the linear optimization in Line 3 of Algorithm 9 or (4.9) in Algorithm 10 can be evaluated simply as $\boldsymbol{a}_t = -r \cdot \mathrm{sign}([\nabla F_t(\boldsymbol{\theta}_t)]_i) \cdot \boldsymbol{e}_i$, where $i = \arg\max_{j\in[n]} |[\nabla F_t(\boldsymbol{\theta}_t)]_j|$. We can derive the following $\mathcal{O}(\sqrt{\log t/t})$ bound for the gradient error:

Proposition 10. *Assume $\|\boldsymbol{A}_t^\top \boldsymbol{A}_t - \mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}]\|_{max} \leq B_1$ and $\|\boldsymbol{A}_t\|_{max} \leq B_2$ almost surely, with $\|\cdot\|_{max}$ being the matrix max norm. Define $c := \max_{\boldsymbol{\theta}\in\mathcal{C}} \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_1$. With probability at least $1 - (1 + 1/n)(\pi^2\epsilon/6)$, the following holds for all $\boldsymbol{\theta} \in \mathcal{C}$ and all $t \geq 1$:*

$$\|\nabla F_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|_{\infty} \leq (cB_1 + \sqrt{mB_2\sigma_w^2})\sqrt{\frac{2(\log(2n^2t^2) - \log\epsilon)}{t}}\ , \quad (4.37)$$

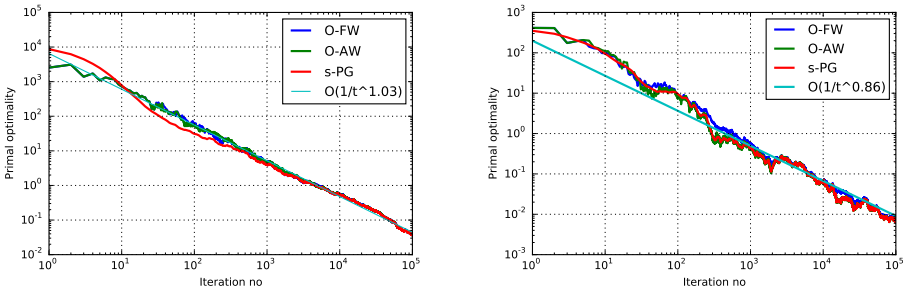

FIGURE 4.1: Online LASSO with synthetic data. Convergence of the primal optimality for online LASSO with (Left) $r = 1.1\|\bar{\boldsymbol{\theta}}\|_1 > \|\boldsymbol{\theta}^\star\|_1$; (Right) $r = 0.15\|\bar{\boldsymbol{\theta}}\|_1 = \|\boldsymbol{\theta}^\star\|_1$.

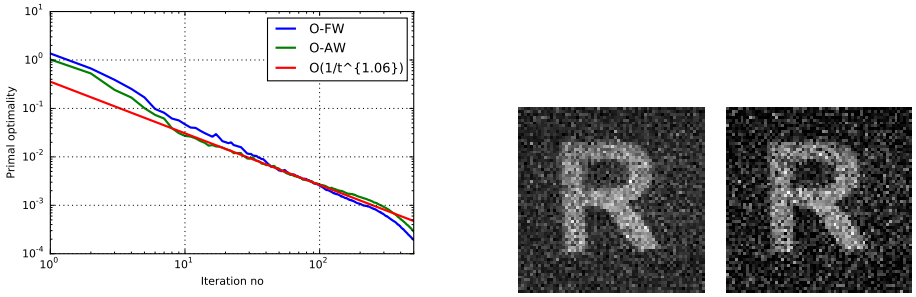

FIGURE 4.2: Online LASSO with single-pixel imaging data `R64.mat`. (Left) Convergence of the objective value. (Middle) Reconstructed image after 500 iterations of O-FW; (Right) O-AW.

where $\|\cdot\|_\infty$ *is the infinity norm and the dual norm of* $\|\cdot\|_1$.

The proof can be founded in subsection 4.6.5. Here, we observe that H11 is satisfied with η_t^ϵ asymptotically equivalent to $4\log(t)$ and $\alpha = 0.5$. Furthermore, the stochastic cost f is L-Lipschitz if $L\mathbf{I} \succeq \mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}]$; μ-strongly convex if $\mathbb{E}[\boldsymbol{A}^\top \boldsymbol{A}] \succeq \mu\mathbf{1}$ for some $\mu > 0$; and H13 is satisfied as $\mathcal{C}$ is a polytope. The analysis from the previous section applies, i.e., O-FW/O-AW has a regret bound of $\mathcal{O}(\log^2 T/T)$.

Synthetic Data. We set $\boldsymbol{A}_t = \boldsymbol{A}$ as fixed for all t with dimension 80×300 and the parameter $\bar{\boldsymbol{\theta}} \in \mathbb{R}^{300}$ is a vector with 10% sparsity and independent $\mathcal{N}(0,1)$ elements. We also set $\sigma_w = 10$. The matrix $\boldsymbol{A}$ is generated as a random Gaussian matrix with independent $\mathcal{N}(0,1)$ elements. For benchmarking purpose, we have compared the O-FW/O-AW's performance with a stochastic projected gradient (sPG) method Rosasco et al. (2014) with a fixed step size $1/L$.

4.1 plots the primal optimality $h_t := f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star)$ with the round number t. The left figure corresponds to the scenario under H12 as $\boldsymbol{\theta}^\star$ belongs to the interior of $\mathcal{C}$. The simulation result corroborates with our analysis, which indicate a fast convergence rate of $\mathcal{O}(1/t)$. In the right figure, we observe that although H12 is not satisfied, the O-FW algorithm still maintains a convergence rate of $\sim \mathcal{O}(1/t)$, and O-AW is slightly outperforming O-FW. Examining the necessity of including H12 in achieving a fast convergence rate for O-FW will be left for future investigation. Lastly, the primal convergence rate of sPG is similar to O-FW. However, the per-iteration complexity of sPG is $\mathcal{O}(n \log n)$, while it is $\mathcal{O}(n)$ for the O-FW.

Next, we consider learning a sparse image $\boldsymbol{\theta}$ from the dataset `R64.mat` available from Duarte et al. (2008). The dataset consists of $T = 4319$ one-bit measurements of a greyscale image of

'R' with size 64×64. The squared loss function is chosen such that $f_t(\boldsymbol{\theta}) = (y_t - \boldsymbol{a}_t^\top \boldsymbol{\theta})^2$, where $\boldsymbol{a}_t \in \mathbb{R}^n$ is a binary measurement vector and $n = 4096$ is the vectorized image. For the O-FW/O-AW algorithms, we have (i) used batch processing by drawing a batch of $B = 5$ new observations and (ii) introduced an inner loop by repeating the O-FW/O-AW updates for 50 times per iteration.

As the optimal solution $\boldsymbol{\theta}^\star$ is unavailable for this problem, Fig. 4.2 compares the primal objective value $F_T(\boldsymbol{\theta}_t)$ against the iteration number and the reconstructed image after $t_f = 500$ iterations of the tested algorithms. The figure shows that the convergence rates of these algorithms all converge at a rate of $\sim \mathcal{O}(1/t)$.

4.5.2 Example: Online matrix completion (MC)

Consider the following setting: we are sequentially given observations in the form (k_t, l_t, Y_t), with $(k_t, l_t) \in [m_1] \times [m_2]$ and $Y_t \in \mathbb{R}$. The observations are assumed to be i.i.d. To define the loss function, the conditional distribution of Y_t *w.r.t.* the sampling is parametrized by an unknown matrix $\bar{X} \in \mathbb{R}^{m_1 \times m_2}$ and supposed to belong to the exponential family, i.e.,

$$p_{\bar{X}}(Y_t | k_t, l_t) := \mathrm{m}(Y_t) \exp\left(Y_t \bar{X}_{k_t,l_t} - \mathrm{A}(\bar{X}_{k_t,l_t})\right), \tag{4.38}$$

where $\mathrm{m}(\cdot)$ and $\mathrm{A}(\cdot)$ are the base measure and log-partition functions, respectively. A natural choice for the loss function at round t is obtained by taking the logarithm of the posterior, *i.e.,*

$$f_t(\boldsymbol{\theta}) := \mathrm{A}(\boldsymbol{\theta}_{k_t,l_t}) - Y_t \boldsymbol{\theta}_{k_t,l_t}.$$

Our goal is to minimize the regret with a penalty favoring low rank solutions $\mathcal{C} := \{\boldsymbol{\theta} \in \mathbb{R}^{m_1 \times m_2} : \|\boldsymbol{\theta}\|_{\sigma,1} \leq R\}$, and the stochastic cost associated is $f(\boldsymbol{\theta}) := \mathbb{E}_{\bar{\boldsymbol{\theta}}}\big[\mathrm{A}(\boldsymbol{\theta}_{k_1,l_1}) - Y_1 \boldsymbol{\theta}_{k_1,l_1}\big]$.

Note that the aggregated gradient $\nabla F_t(\boldsymbol{\theta}_t) = t^{-1} \nabla \sum_{s=1}^t f_s(\boldsymbol{\theta}_t)$ is:

$$\big[\nabla F_t(\boldsymbol{\theta}_t)\big]_{k,l} = t^{-1} \mathrm{A}'([\boldsymbol{\theta}_t]_{k,l}) \big[\textstyle\sum_{s=1}^t e_{k_s} e_{l_s}^{\prime\top}\big]_{k,l} - t^{-1} \big[\textstyle\sum_{s=1}^t Y_s e_{k_s} e_{l_s}^{\prime\top}\big]_{k,l},$$

for all $k, l \in [m_1] \times [m_2]$, with $\{e_k\}_{k=1}^{m_1}$ (*resp.* $\{e'_l\}_{l=1}^{m_2}$) the canonical basis of of $\mathbb{R}^{m_1}$ (*resp.* $\mathbb{R}^{m_2}$). We observe that the two matrices $\sum_{s=1}^t e_{k_s} e_{l_s}^{\prime\top}$ and $\sum_{s=1}^t Y_s e_{k_s} e_{l_s}^{\prime\top}$ can be computed 'on-the-fly' as the running sum. The two matrices can also be stored efficiently in the memory as they are at most t-sparse. The per iteration complexity is upper bounded by $\mathcal{O}(\min\{m_1 m_2, T\})$, where T is the total number of observations.

We observe that for online MC, a better anytime/regret bound than the general case analyzed in Section 4.3 can be achieved. Here, let us state the following assumptions on the observation model:

H1. *The noise variance is finite, that is there exists a constant $\bar{\sigma} > 0$ such that for all $\vartheta \in \mathbb{R}$, $0 \leq \mathrm{A}''(\vartheta) \leq \bar{\sigma}^2$, and the noise is sub-exponential* i.e., *there exist a constant $\lambda \geq 1$ such that for all $(k, l) \in [m_1] \times [m_2]$:*

$$\int \exp\left(\lambda^{-1} \left|y - \mathrm{A}'(\bar{X}_{k,l})\right|\right) p_{\bar{X}}(y|k,l) dy \leq \mathrm{e}\,, \tag{4.39}$$

where $p_{\bar{X}}(\cdot)$ is defined as $p_{\bar{X}}(y|k,l) := \mathrm{m}(y) \exp\left(y \bar{X}_{k,l} - \mathrm{A}(\bar{X}_{k,l})\right)$ and e is the natural number.

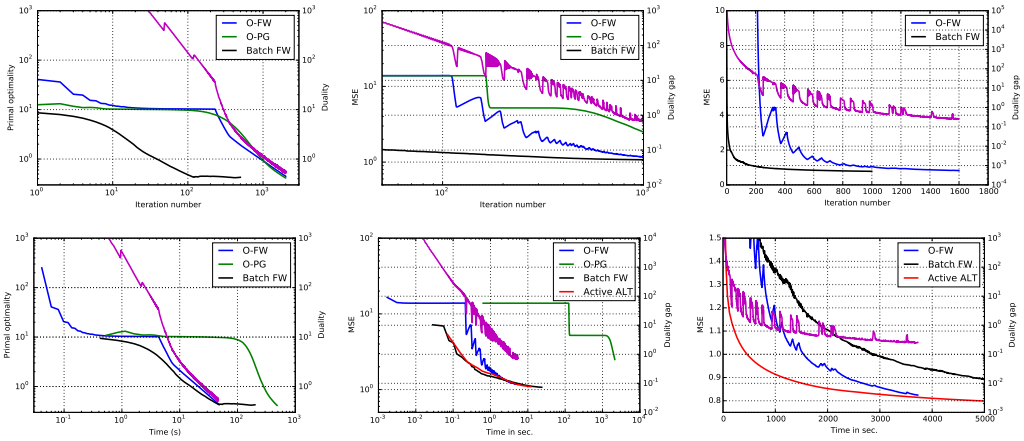

FIGURE 4.3: Online MC performance. (Left) synthetic with batch size $B = 1000$; (Middle) `movielens100k` with $B = 80$; (Right) `movielens20m` with $B = 10000$. (Top) objective value/MSE against round number; (Bottom) against execution time. The duality gap g_t^{FW} is plotted in purple.

H2. *There exists a finite constant $\kappa > 0$ such that for all $\boldsymbol{\theta} \in \mathcal{C}$, $k \in [m_1]$, $l \in [m_2]$*

$$\kappa \geq \max \left(\sqrt{\sum_{l=1}^{m_2} \mathrm{A}'(\boldsymbol{\theta}_{k,l})^2}, \sqrt{\sum_{k=1}^{m_1} \mathrm{A}'(\boldsymbol{\theta}_{k,l})^2} \right) . \tag{4.40}$$

Notice that $\kappa = \mathcal{O}(\sqrt{\max\{m_1, m_2\}})$.

Note that A1 and A2 are satisfied by all exponential family distributions. We can prove the following error bound for the gradient:

Proposition 11. *Assume A1, A2 and that the sampling distribution is uniform. With probability at least $1 - \epsilon$, for any $t \geq T_\epsilon := (\lambda/\bar{\sigma})^2 \log^2(\lambda/\bar{\sigma}) \log(d + 2d/\epsilon)$, and any $\boldsymbol{\theta} \in \mathcal{C}$:*

$$\|\nabla F_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|_{\sigma,\infty} = \mathcal{O}\left(c_\lambda \left(\kappa + \bar{\sigma}\right) \sqrt{\frac{\log(d(1 + t^2/\epsilon))}{t(m_1 \wedge m_2)}} \right) ,$$

with $\|\cdot\|_{\sigma,\infty}$ the operator norm, c_λ a constant which depends only on λ. The constants λ, $\bar{\sigma}$ and κ are defined in A1 and A2.

The proof is relegated to subsection 4.6.6. Armed with the proposition above, we see that the online gradient satisfies H11 with $\eta_t^\epsilon = \mathcal{O}(\log t)$ and $\alpha = 0.5$. Moreover, $f(\boldsymbol{\theta})$ is strongly convex if $\mathrm{A}''(\theta) \geq \mu$. For example, this holds for square loss function. Now if H12 is also satisfied, repeating the analysis in Section 4.3 yields an anytime and regret bound of $\mathcal{O}(\log t/t)$ and $\mathcal{O}(\log^2 T/T)$, respectively.

We test our online MC algorithm on a small synthetically generated dataset, where $\bar{\boldsymbol{\theta}}$ is a rank-20, 200×5000 matrix with Gaussian singular vectors. There are 2×10^6 observations with Gaussian noise of variance 3. Also, we test with two dataset `movielens100k`, `movielens20m` from Harper & Konstan (2015), which contains 10^5, 2×10^7 movie ratings from 943, 138493 users on 1682, 26744 movies, respectively. We assume Gaussian observation and the loss function $f_t(\cdot)$ is designed as the square loss.

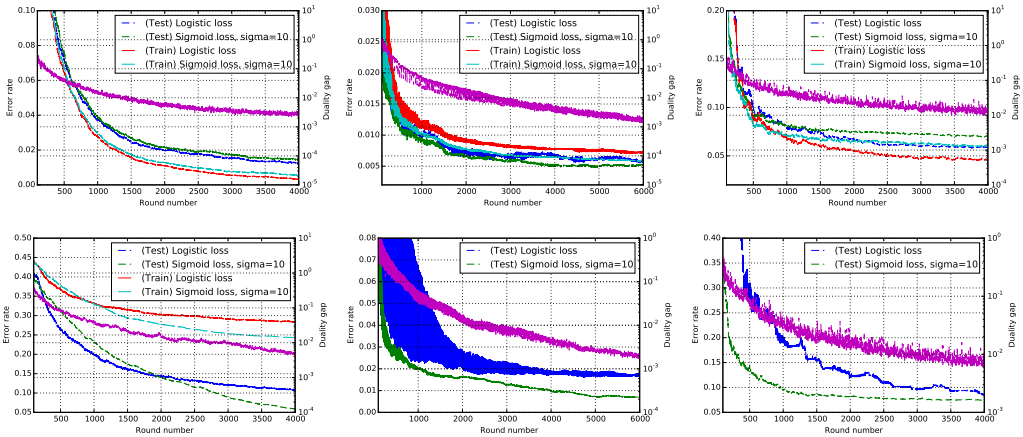

FIGURE 4.4: Binary classification performance against round number t for: (Left) synthetic data; (Middle) `mnist` (class '1'); (Right) `rcv1.binary`. (Top) with no flip (Bottom) with 25% flip in the training labels. The duality gap g_t^{FW} for O-FW with sigmoid loss is plotted in purple.

Results. We compare O-FW to a simple online projected-gradient (O-PG) method. The step size for O-FW is set as $\gamma_t = 2/(1+t)$. For the movielens datasets, the parameter $\bar{\boldsymbol{\theta}}$ is unknown, therefore we split the dataset into training (80%) and testing (20%) set and evaluate the mean square error on the test set. Radiuses of $\mathcal{C}_R$ are set as $R = 1.1\|\bar{\boldsymbol{\theta}}\|_{\sigma,1}$ (synthetic), $R = 10000$ (`movielens100k`) and $R = 150000$ (`movielens20m`). Note that H12 is satisfied by the synthetic case.

The results are shown in Figure 4.3. For the synthetic data, we observe that the stochastic objective of O-FW decreases at a rate $\sim \mathcal{O}(1/t)$, as predicted in our analysis. Significant complexity reduction compared to O-PG for synthetic and `movielens100k` datasets are also observed. The running time is faster than the *batch* FW with line searched step size on `movielens20m`, which we suspect is caused by the simpler linear optimization (4.7) solved at the algorithm initialization by O-FW[1]; and is also comparable to a state-of-the-art, specialized *batch* algorithm for MC problems in Hsieh & Olsen (2014) ('active ALT') and achieves the same MSE level, even though the data are acquired in an online fashion in O-FW.

4.5.3 Example: Robust Binary Classification with Outliers

Consider the following online learning setting: the training data is given sequentially in the form of $(y_t, \boldsymbol{x}_t)$, where $y_t \in \{\pm 1\}$ is a binary label and $\boldsymbol{x}_t \in \mathbb{R}^n$ is a feature vector. Our goal is to train a classifier $\boldsymbol{\theta} \in \mathbb{R}^n$ such that for an arbitrary feature vector $\hat{\boldsymbol{x}}$ it assigns $\hat{y} = \operatorname{sign}(\langle \boldsymbol{\theta}, \hat{\boldsymbol{x}} \rangle)$.

The dataset may sometimes be contaminated by wrong labels. As a remedy, we design a *sigmoid* loss function $f_t(\boldsymbol{\theta}) := (1 + \exp(10 \cdot y_t \langle \boldsymbol{\theta}, \boldsymbol{x}_t \rangle))^{-1}$ that approximates the $0/1$ loss function Shalev-Shwartz et al. (2011); Ertekin et al. (2011). Note that $f_t(\boldsymbol{\theta})$ is smooth and Lipschitz, but *not convex*. For $\mathcal{C}$, we consider the ℓ_1 ball $\mathcal{C}_{\ell_1} = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_1 \leq r\}$ when a sparse classifier is preferred; or the trace-norm ball $\mathcal{C}_\sigma = \{\boldsymbol{\theta} \in \mathbb{R}^{m_1 \times m_2} : \|\boldsymbol{\theta}\|_{\sigma,1} \leq R\}$, where $n = m_1 m_2$, when a low rank classifier is preferred.

We evaluate the performance of our online classifier on synthetic and real data. For the synthetic data, the true classifier $\bar{\boldsymbol{\theta}}$ is a rank-10, 30×30 Gaussian matrix. Each feature $\boldsymbol{x}_t$ is a

[1]This operation amounts to finding the top singular vectors of $\nabla F_t(\boldsymbol{\theta}_t)$, whose complexity grows linearly with the number of non-zeros in $\nabla F_t(\boldsymbol{\theta}_t)$.

30×30 Gaussian matrix. We have 40000 (20000) tuples of data for training (testing). We also test the classifier on the `mnist` (classifying '1' from the rest of the digits), `rcv1.binary` dataset from `LIBSVM` Chang & Lin (2011). The feature dimensions are 784, 47236, and there are 60000 (10000) and 20242 (677399) data tuples for training (testing), respectively. We artificially and randomly flip 0%, 25% labels in the training set.

Results. As benchmark, we compare with the logistic loss function, i.e., $f_t(\boldsymbol{\theta}) = \log(1 + \exp(-y_t\langle\boldsymbol{\theta}, \boldsymbol{x}_t\rangle))$. We apply O-FW with a learning rate of $\alpha = 0.75$ for both loss functions[2], i.e., $\gamma_t = 1/t^{0.75}$. For the synthetic data and `mnist`, the sigmoid (logistic) loss classifier is trained with a trace norm ball constraint of $R = 1$ ($R = 10$). Each round is fed with a batch of $B = 10$ tuples of data. For `rcv1.binary`, we train the classifiers with ℓ_1-ball constraint of $r = 100$ ($r = 1000$) for sigmoid (logistic) loss. Each round is fed with a batch of $B = 5$ tuples of data.

As seen in Figure 4.4, the logistic loss and sigmoid loss performs similarly when there are no flip in the labels; and the sigmoid loss demonstrates better classification performance when some of the labels are flipped. Lastly, the duality gap of O-FW applied to the non-convex loss decays gradually with t, indicating that the algorithm converges to a stationary point.

4.6 Proofs

4.6.1 Proof of Theorem 3

We analyze a slightly more general case where the step size is set as $\gamma_t = (K/K + t - 1)$ for some $K \in \mathbb{Z}_+^\star$. Define $h_t = f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star)$.

4.6.1.1 Slow Convergence of S-FW

For S-FW, we show that with probability at least $1 - \epsilon$:

$$h_t \leq D_0 \left(\frac{\eta_t^\epsilon}{t + K - 1}\right)^\alpha, \ \forall\, t \geq 2, \text{ where } D_0 = \frac{K^2 L\bar{\rho}^2/2 + \rho\sigma K}{K - \alpha} \ . \tag{4.41}$$

Using the L-smoothness of f and the boundedness of $\mathcal{C}$, we get

$$h_{t+1} \leq h_t + \gamma_t\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle + \frac{1}{2}\gamma_t^2 L\bar{\rho}^2 \ . \tag{4.42}$$

On the other hand, the following also hods:

$$\begin{aligned}\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle &= \langle\hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle - \langle\boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle \\ &\leq \langle\hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t\rangle - \langle\boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle \\ &= \langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t\rangle + \langle\boldsymbol{\epsilon}_t, \boldsymbol{\theta}^\star - \boldsymbol{a}_t\rangle \leq -h_t + \rho\|\boldsymbol{\epsilon}_t\| \ . \end{aligned} \tag{4.43}$$

where the second line follows from the definition of $\boldsymbol{a}_t$ and the last inequality is due to the convexity of f and the definition of the diameter. Plugging (4.43) into (4.42) and using H11 yields the following with probability at least $1 - \Delta$ and for all $t \geq 1$

$$h_{t+1} \leq (1 - \gamma_t)h_t + \gamma_t\rho\sigma(\frac{\eta_t^\epsilon}{K + t - 1})^\alpha + \frac{1}{2}\gamma_t^2 L\bar{\rho}^2 \ . \tag{4.44}$$

[2]We have implemented an online variant of projected gradient and observed similar performance as O-FW.

When $t = 2$, the bound (4.41) obviously holds. We now proceed by induction to prove the first bound of the Theorem. Define

$$D_0 = (K^2 L\bar{\rho}^2/2 + \rho\sigma K)/(K - \alpha) \; .$$

The initialization is done by applying (4.44) with $t = 1$ and noting that $K \geq 1$. Assume that $h_t \leq D_0(\eta_t^\epsilon/(K + t - 1))^\alpha$ for some $t \geq 1$. Since $\gamma_t = K/(t + K - 1)$, from (4.44) we get:

$$\begin{aligned} & h_{t+1} - D_0\Big(\frac{\eta_{t+1}^\epsilon}{K+t}\Big)^\alpha \qquad (4.45) \\ & \leq D_0\Big(\big(\frac{\eta_t^\epsilon}{t+K-1}\big)^\alpha - \big(\frac{\eta_{t+1}^\epsilon}{t+K}\big)^\alpha\Big) + \frac{K^2 L\bar{\rho}^2/2 + \rho\sigma K(\eta_t^\epsilon)^\alpha - D_0 K(\eta_t^\epsilon)^\alpha}{(t+K-1)^{1+\alpha}} \\ & \leq (\eta_t^\epsilon)^\alpha\Big(\frac{D_0}{(t+K-1)^\alpha} - \frac{D_0}{(t+K)^\alpha} + \frac{K^2 L\bar{\rho}^2/2 + \rho\sigma K - D_0 K}{(t+K-1)^{1+\alpha}}\Big) \\ & \leq \frac{(\eta_t^\epsilon)^\alpha}{(t+K-1)^{1+\alpha}}\Big((\alpha - K)D_0 + K^2 L\bar{\rho}^2/2 + \rho\sigma K\Big) \leq 0 \; , \end{aligned}$$

where we used the fact that η_t^ϵ is increasing and larger that 1 for the second inequality and $1/(t + K - 1)^\alpha - 1/(t + K)^\alpha \leq \alpha/(t + K - 1)^{1+\alpha}$ for the third inequality. The induction argument is now completed.

4.6.1.2 Slow Convergence of S-AW

For S-AW algorithm, we show that the following holds with probability at least $1 - \epsilon$:

$$h_t \leq D_2'\Big(\frac{\eta_t^\epsilon}{n_{t-1}+K}\Big)^\alpha, \; \forall\, t \geq 2, \text{ where } D_2' = \frac{K}{K-\alpha}(KL\bar{\rho}^2/2 + 2\rho\sigma) \; , \qquad (4.46)$$

Notice that as $n_t \geq t/2$ (cf. Lemma 2), plugging in $K = 2$ yields the desirable result.

The initialization of the induction is easily checked for $t = 2$. We proceed by induction and assume for some $t > 0$ that $h_t \leq D_2'(\eta_t^\epsilon/(n_{t-1} + K))^\alpha$ holds. First of all, observe that from the L-smoothness of $f(\boldsymbol{\theta})$,

$$h_{t+1} \leq h_t + \hat{\gamma}_t\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{d}_t\rangle + \frac{1}{2}\hat{\gamma}_t^2 L\bar{\rho}^2 \; . \qquad (4.47)$$

Moreover, we have:

$$\begin{aligned} \langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{d}_t\rangle &= \langle\hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{d}_t\rangle - \langle\boldsymbol{\epsilon}_t, \boldsymbol{d}_t\rangle \leq \langle\hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t\rangle - \langle\boldsymbol{\epsilon}_t, \boldsymbol{d}_t\rangle \\ &\leq \langle\hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_\star - \boldsymbol{\theta}_t\rangle - \langle\boldsymbol{\epsilon}_t, \boldsymbol{d}_t\rangle = \langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_\star - \boldsymbol{\theta}_t\rangle + \langle\boldsymbol{\epsilon}_t, \boldsymbol{\theta}_\star - \boldsymbol{\theta}_t - \boldsymbol{d}_t\rangle \\ &\leq -h_t + 2\rho\|\boldsymbol{\epsilon}_t\| \qquad (4.48) \end{aligned}$$

where we used the condition of line 4 (Algorithm 10) in the first inequality and the fact $\|\boldsymbol{\theta}_\star - \boldsymbol{\theta}_t - \boldsymbol{d}_t\|_\star \leq 2\rho$ in the last inequality. This gives

$$h_{t+1} \leq (1 - \hat{\gamma}_t)h_t + 2\hat{\gamma}_t\rho\sigma\Big(\frac{\eta_t^\epsilon}{K+n_{t-1}}\Big)^\alpha + \frac{1}{2}\hat{\gamma}_t^2 L\bar{\rho}^2 \; , \qquad (4.49)$$

where we have used H11 and the fact that $n_{t-1} \leq t - 1$.

Consider the two cases: when a drop step (line 11) is taken at iteration $t + 1$, the following result gives the induction.

Lemma 12. *Suppose that* $h_t \leq D_2'(\eta_t^\epsilon/(K+n_{t-1}))^{2\alpha}$ *for* $\alpha \in (0,1]$*, and that a drop step is taken at time* $t+1$ *(see Algorithm 10 line 11), then*

$$h_{t+1} \leq D_2' \left(\frac{\eta_{t+1}^\epsilon}{K+n_t} \right)^\alpha . \tag{4.50}$$

The proof can be found at the end of this section. On the other hand, when a drop step is *not* taken, notice that we will have $\hat{\gamma}_t = \gamma_{n_t} = K/(K+n_t-1)$ and $n_t = n_{t-1}+1$.

Consequently, the same induction argument for S-FW (replacing t by n_t and consider $h_{t+1} - D_2'(\eta_{t+1}^\epsilon/(K+n_t))^\alpha$) shows that $h_{t+1} \leq D_2' \left(\frac{\eta_{t+1}^\epsilon}{K+n_t} \right)^\alpha$. Obviously, the induction step is completed since both cases yield the desirable bound.

Proof of Lemma 12. Using (4.49) gives the following chain

$$\begin{aligned}
&h_{t+1} - D_2' \left(\frac{\eta_{t+1}^\epsilon}{K+n_t} \right)^\alpha \\
&\quad \leq (1-\hat{\gamma}_t) h_t + 2\hat{\gamma}_t \rho\sigma \left(\frac{\eta_{t+1}^\epsilon}{K+n_t} \right)^\alpha + \frac{1}{2} L\bar{\rho}^2 \hat{\gamma}_t^2 - D_2' \left(\frac{\eta_{t+1}^\epsilon}{K+n_t} \right)^\alpha \\
&\quad \leq (1-\hat{\gamma}_t) D_2' \left(\frac{\eta_t^\epsilon}{K+n_t} \right)^\alpha + 2\hat{\gamma}_t \rho\sigma \left(\frac{\eta_{t+1}^\epsilon}{K+n_t} \right)^\alpha + \frac{1}{2} L\bar{\rho}^2 \hat{\gamma}_t^2 - D_2' \left(\frac{\eta_{t+1}^\epsilon}{K+n_t} \right)^\alpha \\
&\quad \leq \hat{\gamma}_t \left((-D_2' + 2\rho\sigma) \left(\frac{\eta_t^\epsilon}{K+n_t} \right)^\alpha + \hat{\gamma}_t \frac{L\bar{\rho}^2}{2} \right) \\
&\quad \leq \hat{\gamma}_t \left(\left(-D_2' + 2\rho\sigma + \frac{1}{2} KL\bar{\rho}^2 \right) \left(\frac{\eta_t^\epsilon}{K+n_t} \right)^\alpha \right) \leq 0 .
\end{aligned}$$

In the above, the second inequality is due to $1-\hat{\gamma}_t \geq 0$ and the induction hypothesis; the third inequality is due to η_t^ϵ is increasing and; the last inequality is due to $\hat{\gamma}_t < K/(K+n_t)$. The proof is completed.

4.6.2 Proof of Theorem 4

Like the proof for Theorem 3, the analysis below is done by assuming a more general step size rule $\gamma_t = K/(K+t-1)$ with some $K \in \mathbb{Z}_+^\star$. Let $h_t := f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}^\star)$. As explained in the proof sketch, we first state the following lemma:

Lemma 13. *Lacoste-Julien & Jaggi (2013, 2015) Consider Algorithm 9. Assume H12 and that* f *is* L*-smooth and* μ*-strongly convex, then*

$$\big(\max_{\boldsymbol{\theta} \in \mathcal{C}} \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta} \rangle \big)^2 \geq 2\mu\delta^2 h_t \quad \text{and} \quad L\bar{\rho}^2 \geq \mu\delta^2 . \tag{4.51}$$

Consider Algorithm 10, assume H13 and that f *is* L*-smooth and* μ*-strongly convex, then*

$$\big(\max_{\boldsymbol{\theta} \in \mathcal{A}_t} \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta} \rangle - \min_{\boldsymbol{\theta} \in \mathcal{C}} \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta} \rangle \big)^2 \geq 2\mu\delta_{\text{AW}}^2 h_t \quad \text{and} \quad L\bar{\rho}^2 \geq \mu\delta_{\text{AW}}^2 . \tag{4.52}$$

For completeness, the proof can be found in Section 4.6.2.3. We remark that the above lemma is a key result leading to the linear convergence of the classical FW/AW algorithms with *adaptive* step sizes, as studied in Lacoste-Julien & Jaggi (2013, 2015). As we shall show below, Lemma 13 enables us to prove Theorem 4 for the accelerated convergence rate of S-FW/S-AW.

4.6.2.1 Fast Convergence of S-FW

Define $\beta = 1 + 2\alpha/(K-\alpha)$. For the S-FW, we show that with probability at least $1-\epsilon$ and for all $t \geq 2$,

$$h_t \leq D_1 \left(\frac{\eta_t^\epsilon}{t+K-1} \right)^{2\alpha} , \text{ where } D_1 = \max\left\{ 4\Big(\frac{K+1}{K}\Big)^{2\alpha}, \beta^2 \right\} \frac{(\rho\sigma + KL\bar{\rho}^2/2)^2}{2\delta^2\mu} . \tag{4.53}$$

Define $\boldsymbol{\epsilon}_t = \hat{\nabla}_t f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)$ and recall that $g_t = \max_{\boldsymbol{s}\in\mathcal{C}} \langle \boldsymbol{\theta}_t - \boldsymbol{s}, \nabla f(\boldsymbol{\theta}_t)\rangle$ is the FW gap at $\boldsymbol{\theta}_t$. Notice that (4.51) in Lemma 13 implies $g_t \geq \sqrt{2\mu\delta^2 h_t}$. Furthermore, we define $\boldsymbol{s}_t \in \arg\max_{\boldsymbol{s}\in\mathcal{C}} \langle \boldsymbol{\theta}_t - \boldsymbol{s}, \nabla f(\boldsymbol{\theta}_t)\rangle$. We note that

$$\begin{aligned} \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle &\leq \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{s}_t - \boldsymbol{\theta}_t \rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{a}_t - \boldsymbol{\theta}_t \rangle & (4.54) \\ &= \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{s}_t - \boldsymbol{\theta}_t \rangle + \langle \boldsymbol{\epsilon}_t, \boldsymbol{s}_t - \boldsymbol{a}_t \rangle \\ &\leq -g_t + \rho\|\boldsymbol{\epsilon}_t\| \leq -\delta\sqrt{2\mu h_t} + \rho\|\boldsymbol{\epsilon}_t\| , & (4.55) \end{aligned}$$

where the last line follows from Lemma 13. Combining the L-smoothness of $f(\boldsymbol{\theta})$ and (4.55) yield the following with probability at least $1-\epsilon$ and for all $t \geq 1$,

$$h_{t+1} \leq \sqrt{h_t}(\sqrt{h_t} - \gamma_t\delta\sqrt{2\mu}) + \gamma_t\rho\sigma \left(\frac{\eta_t^\epsilon}{t+K-1} \right)^{\alpha} + \frac{1}{2}\gamma_t^2 L\bar{\rho}^2 . \tag{4.56}$$

Let us proceed by induction. Suppose that $h_t \leq D_1(\eta_t^\epsilon/(t+K-1))^{2\alpha}$ for some $t \geq 1$. There are two cases to analyze.

Case 1: when $h_t - \gamma_t\delta\sqrt{2\mu h_t} \leq 0$ — Since $\gamma_t = K/(K+t-1)$, (4.56) yields

$$\begin{aligned} h_{t+1} &\leq \rho\sigma K \frac{(\eta_t^\epsilon)^\alpha}{(K+t-1)^{1+\alpha}} + \frac{L\bar{\rho}^2K^2}{2(K+t-1)^2} \leq (\rho\sigma K + L\bar{\rho}^2K^2/2)\frac{(\eta_{t+1}^\epsilon)^{2\alpha}}{(K+t-1)^{2\alpha}} \\ &\leq (\rho\sigma K + L\bar{\rho}^2K^2/2)\Big(\frac{K+1}{K}\Big)^{2\alpha}\Big(\frac{\eta_{t+1}^\epsilon}{K+t}\Big)^{2\alpha} , \end{aligned}$$

where we used the fact that η_t^ϵ is increasing and larger than 1. To conclude, one just needs to check that

$$(\rho\sigma K + L\bar{\rho}^2K^2/2)\Big(\frac{K+1}{K}\Big)^{2\alpha} \leq D_1 . \tag{4.57}$$

However, we note that

$$D_1 \geq \Big(\frac{K+1}{K}\Big)^{2\alpha}(\rho\sigma + L\bar{\rho}^2K/2)\frac{(\rho\sigma + L\bar{\rho}^2K/2)}{\mu\delta^2/2} \geq \Big(\frac{K+1}{K}\Big)^{2\alpha}(\rho\sigma + L\bar{\rho}^2K/2)K ,$$

where the last inequality is due to $L\bar{\rho}^2 \geq \delta^2\mu$ from Lemma 13. Hence, we obtain $h_{t+1} \leq D_1(\eta_{t+1}^\epsilon/(K+t))^{2\alpha}$ and the induction step is completed.

Case 2: when $h_t - \gamma_t \delta \sqrt{2\mu h_t} > 0$ — By induction hypothesis and (4.56), we have

$$\begin{aligned}
&h_{t+1} - D_1\Big(\frac{\eta_{t+1}^\epsilon}{K+t}\Big)^{2\alpha} \\
&\le D_1\Big(\Big(\frac{\eta_t^\epsilon}{K+t-1}\Big)^{2\alpha} - \Big(\frac{\eta_{t+1}^\epsilon}{K+t}\Big)^{2\alpha}\Big) + \frac{(\eta_t^\epsilon)^\alpha \cdot K}{(K+t-1)^{1+\alpha}}\Big(\rho\sigma + \frac{L\bar{\rho}^2 K}{2} - \delta\sqrt{2\mu D_1}\Big) \\
&\le \frac{(\eta_t^\epsilon)^\alpha}{(K+t-1)^{1+\alpha}}\left[2\alpha D_1\left(\frac{\eta_t^\epsilon}{t+K-1}\right)^\alpha + K\rho\sigma + K^2 L\bar{\rho}^2/2 - \delta K\sqrt{2\mu D_1}\right] \\
&\le \frac{(\eta_t^\epsilon)^\alpha}{(K+t-1)^{1+\alpha}}\left[2\alpha D_1\left(\frac{\eta_t^\epsilon}{t+K-1}\right)^\alpha + (K\rho\sigma + K^2 L\bar{\rho}^2/2)(1-\beta)\right]
\end{aligned}$$

where we used the fact that (i) η_t^ϵ is increasing and larger than 1, (ii) $t \ge 1$ and (iii) $1/(K+t-1)^{2\alpha} - 1/(K+t)^{2\alpha} \le 2\alpha/(K+t-1)^{1+2\alpha}$ in the second last inequality; and we have used the definition of D_1 in the last inequality. Define

$$t_0 := \inf\{t \ge 1 : 2\alpha D_1\left(\frac{\eta_t^\epsilon}{t+K-1}\right)^\alpha + (K\rho\sigma + K^2 L\bar{\rho}^2/2)(1-\beta) \le 0\}. \tag{4.58}$$

Since $\eta_t^\epsilon/(K+t-1)$ is monotonically decreasing to 0 and $\beta > 1$, t_0 exists. Clearly, for any $t > t_0$ the RHS is non-positive. For $t \le t_0$, we have

$$(K\rho\sigma + K^2 L\bar{\rho}^2/2)(\beta-1) \le 2\alpha D_1\left(\frac{\eta_t^\epsilon}{t+K-1}\right)^\alpha \tag{4.59}$$

i.e.,

$$D_0(K-\alpha)(\beta-1) \le 2\alpha D_1\left(\frac{\eta_t^\epsilon}{t+K-1}\right)^\alpha \tag{4.60}$$

Hence by the definition that $\beta = 1 + 2\alpha/(K-\alpha)$ and applying Theorem 3, we get:

$$h_t \le D_0\left(\frac{\eta_t^\epsilon}{t+K-1}\right)^\alpha \le D_1\left(\frac{\eta_t^\epsilon}{t+K-1}\right)^{2\alpha}$$

The initialization is easily verified as the first inequality holds true for all $t \ge 2$.

4.6.2.2 Fast Convergence of S-AW

Define $\beta = 1 + 2\alpha/(K-\alpha)$. We show that with probability at least $1-\epsilon$ and for all $t \ge 2$:

$$h_t \le D_2\Big(\frac{\eta_t^\epsilon}{n_{t-1}+K}\Big)^{2\alpha} \text{ where } D_2 = \max\left\{\Big(\frac{K+1}{K}\Big)^{2\alpha}, \beta^2\right\}\frac{(2\rho\sigma + KL\bar{\rho}^2/2)^2}{(\delta_{\mathrm{AW}})^2\mu/2} \tag{4.61}$$

and n_t is the number of non-drop steps (see Algorithm 10) up to iteration t. Recalling that $n_t \ge t/2$ (cf. Lemma 2), the upper bound above gives the desirable fast convergence rate for S-AW algorithm.

To begin our proof, we define $\boldsymbol{\epsilon}_t = \hat{\nabla}_t f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)$ and the following quantities,

$$\boldsymbol{b}_t^{\mathsf{FW}} := \arg\min_{\boldsymbol{b}\in\mathcal{C}} \left\langle \boldsymbol{b}, \nabla f(\boldsymbol{\theta}_t)\right\rangle,\ \boldsymbol{b}_t^{\mathsf{AW}} := \arg\max_{\boldsymbol{b}\in\mathcal{A}_t} \left\langle \boldsymbol{b}, \nabla f(\boldsymbol{\theta}_t)\right\rangle,$$

$$\bar{g}_t^{\mathsf{AW}} := \left\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{b}_t^{\mathsf{AW}} - \boldsymbol{b}_t^{\mathsf{FW}}\right\rangle. \tag{4.62}$$

We remark that $\boldsymbol{b}_t^{\mathsf{AW}} \neq \boldsymbol{a}_t^{\mathsf{AW}}$ and $\boldsymbol{b}_t^{\mathsf{FW}} \neq \boldsymbol{a}_t^{\mathsf{FW}}$ as the former are evaluated on the true gradient $\nabla f(\boldsymbol{\theta}_t)$. In Algorithm 10, we choose $\boldsymbol{d}_t$ such that $\langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{d}_t \rangle = \min\{\langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t \rangle, \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{a}_t^{\mathsf{AW}} \rangle\}$. Therefore, for any $t \geq 2$:

$$\begin{aligned}\langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{d}_t \rangle &\leq \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \frac{\boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{a}_t^{\mathsf{AW}}}{2} \rangle \leq \langle \hat{\nabla} f(\boldsymbol{\theta}_t), \frac{\boldsymbol{b}_t^{\mathsf{FW}} - \boldsymbol{b}_t^{\mathsf{AW}}}{2} \rangle \\ &= \langle \nabla f(\boldsymbol{\theta}_t), \frac{\boldsymbol{b}_t^{\mathsf{FW}} - \boldsymbol{b}_t^{\mathsf{AW}}}{2} \rangle + \langle \boldsymbol{\epsilon}_t, \frac{\boldsymbol{b}_t^{\mathsf{FW}} - \boldsymbol{b}_t^{\mathsf{AW}}}{2} \rangle\end{aligned}$$

where the second inequality is due to the definitions of $\boldsymbol{a}_t^{\mathsf{FW}}$ and $\boldsymbol{a}_t^{\mathsf{AW}}$ in (4.9). Hence:

$$\langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{d}_t \rangle \leq -\frac{\bar{g}_t^{\mathsf{AW}}}{2} + \langle \boldsymbol{\epsilon}_t, \frac{\boldsymbol{b}_t^{\mathsf{FW}} - \boldsymbol{b}_t^{\mathsf{AW}}}{2} \rangle \,. \tag{4.63}$$

As f is L-smooth, the following holds,

$$\begin{aligned}f(\boldsymbol{\theta}_{t+1}) &\leq f(\boldsymbol{\theta}_t) + \hat{\gamma}_t \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{d}_t \rangle + \frac{L\bar{\rho}^2}{2}\hat{\gamma}_t^2 \\ &= f(\boldsymbol{\theta}_t) + \hat{\gamma}_t(\langle \hat{\nabla} f(\boldsymbol{\theta}_t), \boldsymbol{d}_t \rangle - \langle \boldsymbol{\epsilon}_t, \boldsymbol{d}_t \rangle) + \hat{\gamma}_t^2 \frac{L\bar{\rho}^2}{2} \\ &\leq f(\boldsymbol{\theta}_t) - \hat{\gamma}_t \frac{\bar{g}_t^{\mathsf{AW}}}{2} + \hat{\gamma}_t \langle \boldsymbol{\epsilon}_t, \frac{\boldsymbol{b}_t^{\mathsf{FW}} - \boldsymbol{b}_t^{\mathsf{AW}}}{2} - \boldsymbol{d}_t \rangle + \hat{\gamma}_t^2 \frac{L\bar{\rho}^2}{2} \,,\end{aligned} \tag{4.64}$$

where we used (4.63) for the last line. Using $\|(\boldsymbol{b}_t^{\mathsf{FW}} - \boldsymbol{b}_t^{\mathsf{AW}})/2 - \boldsymbol{d}_t\|_* \leq 2\rho$, subtracting $f(\boldsymbol{\theta}^*)$ on both sides and applying H11 yield

$$h_{t+1} \leq h_t - \hat{\gamma}_t \frac{\bar{g}_t^{\mathsf{AW}}}{2} + 2\hat{\gamma}_t \rho\sigma \left(\frac{\eta_t^\epsilon}{K + t - 1} \right)^\alpha + \hat{\gamma}_t^2 \frac{L\bar{\rho}^2}{2} \,, \tag{4.65}$$

We now proceed by induction and assume that for some $t \geq 2$, $h_t \leq D_2(\eta_t^\epsilon/(K + n_{t-1}))^{2\alpha}$ holds. Notice that (4.52) in Lemma 13 gives $\bar{g}_t^{\mathsf{AW}} \geq \sqrt{2\mu\delta_{\mathrm{AW}}^2 h_t}$. Suppose that $h_t > 0$ ($h_t = 0$ is discussed at the end of the proof). Combining (4.65) and Lemma 13 give:

$$h_{t+1} \leq h_t - \hat{\gamma}_t \delta_{\mathrm{AW}} \sqrt{\frac{\mu h_t}{2}} + 2\hat{\gamma}_t \rho\sigma \Big(\frac{\eta_t^\epsilon}{n_{t-1} + K} \Big)^\alpha + \hat{\gamma}_t^2 \frac{L\bar{\rho}^2}{2} \,. \tag{4.66}$$

Consider two different cases. When a drop step is taken at iteration $t + 1$, the induction step can be taken care of by:

Lemma 14. *Suppose that $h_t \leq D_2(\eta_t^\epsilon/(K+n_{t-1}))^{2\alpha}$ and that a drop step is taken at iteration $t + 1$ (see Algorithm 10 line 11), then*

$$h_{t+1} \leq D_2 \left(\frac{\eta_{t+1}^\epsilon}{K + n_t} \right)^{2\alpha} \,, \tag{4.67}$$

note that $n_t = n_{t-1}$ when a drop step is taken.

The proof can be found at the end of this section. The above lemma shows that our upper bound on the objective value does not increase when a drop step is taken.

On the other hand, when a drop step is *not* taken at iteration $t+1$, then from Algorithm 10, we have $\hat{\gamma}_t = \gamma_{n_t} = K/(K + n_t - 1)$ and $n_t = n_{t-1} + 1$. We consider the following two cases:

Case 1: If $h_t - \hat{\gamma}_t \delta_{\text{AW}} \sqrt{\frac{\mu h_t}{2}} \leq 0$ — Since $\hat{\gamma}_t = K/(K+n_t-1)$, $n_t \leq t$, Eq. (4.66) yields

$$\begin{aligned} h_{t+1} &\leq 2\rho\sigma K \frac{(\eta_t^\epsilon)^\alpha}{(K+n_t-1)^{1+\alpha}} + \frac{L\bar{\rho}^2 K^2}{2(K+n_t-1)^2} \\ &\leq (2\rho\sigma K + L\bar{\rho}^2 K^2/2) \frac{(\eta_{t+1}^\epsilon)^{2\alpha}}{(K+n_t-1)^{2\alpha}} \\ &\leq (2\rho\sigma K + L\bar{\rho}^2 K^2/2) \Big(\frac{K+1}{K}\Big)^{2\alpha} \left(\frac{\eta_{t+1}^\epsilon}{K+n_t}\right)^{2\alpha} , \end{aligned} \tag{4.68}$$

where we used that η_t^ϵ is increasing and larger than 1. One just needs to check that

$$(2\rho\sigma K + L\bar{\rho}^2 K^2/2) \Big(\frac{K+1}{K}\Big)^{2\alpha} \leq D_2 \, . \tag{4.69}$$

However, we observe that

$$D_2 \geq \Big(\frac{K+1}{K}\Big)^{2\alpha} (2\rho\sigma + \frac{L\bar{\rho}^2 K}{2}) \cdot \frac{4\rho\sigma + L\bar{\rho}^2 K}{\mu \delta_{\text{AW}}^2} \geq \Big(\frac{K+1}{K}\Big)^{2\alpha} (2\rho\sigma + \frac{L\bar{\rho}^2 K}{2}) \cdot K \, ,$$

where the last inequality is due to $L\bar{\rho}^2 \geq \delta_{\text{AW}}^2 \mu$ from Lemma 13. Hence, we obtain $h_{t+1} \leq D_2 (\eta_{t+1}^\epsilon/(K+n_t))^{2\alpha}$ and the induction step is completed.

Case 2: If $h_t - \hat{\gamma}_t \delta_{\text{AW}} \sqrt{\frac{\mu h_t}{2}} > 0$ — By induction and (4.66), we have

$$\begin{aligned} & h_{t+1} - D_2 \Big(\frac{\eta_{t+1}^\epsilon}{K+n_t}\Big)^{2\alpha} \\ &\leq D_2 \Big(\Big(\frac{\eta_t^\epsilon}{K+n_t-1}\Big)^{2\alpha} - \Big(\frac{\eta_{t+1}^\epsilon}{K+n_t}\Big)^{2\alpha} \Big) \\ &\qquad + \frac{(\eta_t^\epsilon)^\alpha \cdot K}{(n_t+K-1)^{1+\alpha}} \Big(2\rho\sigma + L\bar{\rho}^2 K/2 - \delta_{\text{AW}} \sqrt{\frac{\mu D_2}{2}} \Big) \\ &\leq \frac{(\eta_t^\epsilon)^\alpha}{(K+n_t-1)^{1+\alpha}} \Big[2\alpha D_2 \left(\frac{\eta_t^\epsilon}{n_t+K-1}\right)^\alpha + 2K\rho\sigma + \frac{K^2 L\bar{\rho}^2}{2} - \delta_{\text{AW}} K \sqrt{\frac{\mu D_2}{2}} \Big] \\ &\leq \frac{(\eta_t^\epsilon)^\alpha}{(K+n_t-1)^{1+\alpha}} \Big[2\alpha D_2 \left(\frac{\eta_t^\epsilon}{n_t+K-1}\right)^\alpha + (2K\rho\sigma + \frac{K^2 L\bar{\rho}^2}{2})(1-\beta) \Big] \end{aligned}$$

where we used the fact that (i) η_t^ϵ is increasing and larger than 1, (ii) $t \geq 1$ and (iii) $1/(K+t-1)^{2\alpha} - 1/(K+t)^{2\alpha} \leq 2\alpha/(K+t-1)^{1+2\alpha}$ in the second last inequality; and we have used the definition of D_2 in the last inequality. Define

$$t_0 := \inf\{t \geq 1 : 2\alpha D_2 \left(\frac{\eta_t^\epsilon}{n_t+K-1}\right)^\alpha + K(2\rho\sigma + KL\bar{\rho}^2/2)(1-\beta) \leq 0\}. \tag{4.70}$$

Since $\eta_t^\epsilon/(K+n_t-1)$ decreases to 0 (see H11 and Lemma 2), t_0 exists. Clearly, for any $t > t_0$ the RHS is non-positive. For $t \leq t_0$, we have

$$K(2\rho\sigma + KL\bar{\rho}^2/2)(\beta-1) \leq 2\alpha D_2 \left(\frac{\eta_t^\epsilon}{n_t+K-1}\right)^\alpha \tag{4.71}$$

implying

$$D_2'(K-\alpha)(\beta-1) \leq 2\alpha D_2 \left(\frac{\eta_t^\epsilon}{n_t+K-1}\right)^\alpha \tag{4.72}$$

Since $\beta = 1 + 2\alpha/(K-\alpha)$, the left hand side of (4.59) equals $2\alpha D_2'$ and we conclude that $D_2' \leq D_2(\eta_t^\epsilon/(n_t+K-1))^\alpha$. Applying Theorem 3 we get:

$$h_t \leq D_2' \left(\frac{\eta_t^\epsilon}{n_t+K-1}\right)^\alpha \leq D_2 \left(\frac{\eta_t^\epsilon}{n_t+K-1}\right)^{2\alpha}$$

The induction step is completed by observing that $n_t - 1 = n_{t-1}$. The initialization is easily verified for $t = 2$. If $h_t = 0$, then by Lemma 13 yields $g_t^{\text{AW}} = 0$ and the induction is treated as **Case 1**.

Proof of Lemma 14. Since iteration $t+1$ is a drop step, by construction,

$$\hat{\gamma}_t = \gamma_{\max} \leq \frac{K}{K+n_t} \quad \text{and} \quad n_t = n_{t-1}\,.$$

From (4.66) and the assumption in the lemma, we consider two cases: if $\sqrt{h_t} - \hat{\gamma}_t\sqrt{\mu\delta_{\text{AW}}^2/2} \leq 0$, then we have

$$\begin{aligned} h_{t+1} - D_2\left(\frac{\eta_{t+1}^\epsilon}{K+n_t}\right)^{2\alpha} &\leq 2\hat{\gamma}_t\rho\sigma\Big(\frac{\eta_t^\epsilon}{n_{t-1}+K}\Big)^\alpha + \frac{L\bar{\rho}^2\hat{\gamma}_t^2}{2} - D_2\left(\frac{\eta_{t+1}^\epsilon}{n_t+K}\right)^{2\alpha} \\ &\leq 2\rho\sigma\frac{K\cdot(\eta_{t+1}^\epsilon)^\alpha}{(n_t+K)^{1+\alpha}} + \frac{L\bar{\rho}^2}{2}\left(\frac{K}{n_t+K}\right)^2 - D_2\left(\frac{\eta_{t+1}^\epsilon}{n_t+K}\right)^{2\alpha} \\ &\leq \left(\frac{\eta_{t+1}^\epsilon}{n_t+K}\right)^{2\alpha}\Big(2\rho\sigma K + K^2L\bar{\rho}^2/2 - D_2\Big) \end{aligned} \tag{4.73}$$

The second inequality is due to $n_t = n_{t-1}$ and $\hat{\gamma}_t = \gamma_{max} \leq K/(K+n_t)$. The last inequality is due to $2\alpha \leq \min\{2, 1+\alpha\}$ for all $\alpha \in (0,1]$ and η_t^ϵ is an increasing sequence with $\eta_t^\epsilon \geq 1$. It can be verified that the right hand side is non-positive using the definition of D_2.

On the other hand, if $\sqrt{h_t} - \hat{\gamma}_t\sqrt{\mu\delta_{\text{AW}}^2/2} > 0$, we have from (4.66)

$$\begin{aligned} &h_{t+1} - D_2\left(\frac{\eta_{t+1}^\epsilon}{n_t+K}\right)^{2\alpha} \\ &\leq \sqrt{h_t}\Big(\sqrt{h_t} - \hat{\gamma}_t\sqrt{\frac{\mu\delta_{\text{AW}}^2}{2}}\Big) + \frac{L\bar{\rho}^2\hat{\gamma}_t^2}{2} + 2\hat{\gamma}_t\rho\sigma\Big(\frac{\eta_t^\epsilon}{n_{t-1}+K}\Big)^\alpha - D_2\left(\frac{\eta_{t+1}^\epsilon}{n_t+K}\right)^{2\alpha} \\ &\leq \frac{L\bar{\rho}^2\hat{\gamma}_t^2}{2} + 2\hat{\gamma}_t\rho\sigma\Big(\frac{\eta_t^\epsilon}{n_{t-1}+K}\Big)^\alpha - \hat{\gamma}_t\sqrt{D_2\frac{\mu\delta_{\text{AW}}^2}{2}}\left(\frac{\eta_t^\epsilon}{n_{t-1}+K}\right)^\alpha \\ &= \hat{\gamma}_t\Big(\frac{L\bar{\rho}^2\hat{\gamma}_t}{2} + 2\rho\sigma\Big(\frac{\eta_t^\epsilon}{n_t+K}\Big)^\alpha - \sqrt{D_2\frac{\mu\delta_{\text{AW}}^2}{2}}\left(\frac{\eta_t^\epsilon}{n_t+K}\right)^\alpha\Big) \\ &\leq \hat{\gamma}_t\Big(\frac{KL\bar{\rho}^2/2}{n_t+K} + \Big(2\rho\sigma - \sqrt{D_2\frac{\mu\delta_{\text{AW}}^2}{2}}\Big)\left(\frac{\eta_t^\epsilon}{n_t+K}\right)^\alpha\Big) \\ &\leq \hat{\gamma}_t\Big(\frac{\eta_t^\epsilon}{n_t+K}\Big)^\alpha\Big(\frac{KL\bar{\rho}^2}{2} + 2\rho\sigma - \sqrt{D_2\frac{\mu\delta_{\text{AW}}^2}{2}}\Big)\,. \end{aligned}$$

The last inequality is due to $\alpha \leq 1$. Similarly, by the definition of D_2, we observe that the RHS in the above inequality is non-positive.

4.6.2.3 Proof of Lemma 13

We first prove the first part of the lemma, i.e., (4.51), pertaining to the O-FW algorithm. Let $\bar{\boldsymbol{s}}_t \in \partial\mathcal{C}$ be a point on the boundary of $\mathcal{C}$ such that it is co-linear with $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}_t$. Moreover, we defin $g_t := \max_{\boldsymbol{\theta}\in\mathcal{C}}\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle$. As $\boldsymbol{\theta}^\star \in \operatorname{int}(\mathcal{C})$, we can write

$$\boldsymbol{\theta}^\star = \boldsymbol{\theta}_t + \bar{\gamma}(\bar{\boldsymbol{s}}_t - \boldsymbol{\theta}_t) \quad \text{for some} \quad \bar{\gamma} \in [0, 1) \ . \tag{4.74}$$

From the μ-strong convexity of f, we have

$$\frac{\mu}{2}\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_t\|_2^2 \leq f(\boldsymbol{\theta}^\star) - f(\boldsymbol{\theta}_t) - \langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t\rangle = -h_t + \bar{\gamma}\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \bar{\boldsymbol{s}}_t\rangle \leq -h_t + \bar{\gamma} g_t \ ,$$

where the last inequality is due to the definition of g_t. Now, the left hand side of the inequality above can be bounded as

$$\frac{\mu}{2}\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_t\|_2^2 = \bar{\gamma}^2\frac{\mu}{2}\|\bar{\boldsymbol{s}}_t - \boldsymbol{\theta}_t\|_2^2 \geq \bar{\gamma}^2\frac{\mu}{2}\|\bar{\boldsymbol{s}}_t - \boldsymbol{\theta}^\star\|_2^2 \geq \bar{\gamma}^2\delta^2\frac{\mu}{2} \tag{4.75}$$

Combining the two inequalities above yields

$$h_t \leq \bar{\gamma} g_t - \bar{\gamma}^2\delta^2\frac{\mu}{2} \leq \frac{g_t^2}{2\delta^2\mu} \ , \tag{4.76}$$

where the upper bound is achieved by setting $\bar{\gamma} = g_t/(\delta^2\mu)$. Recalling the definition of g_t concludes the proof of the first part. Lastly, we note by combining Eq. (2), Remark 1 and Lemma 2 in Lacoste-Julien & Jaggi (2013), we have $L\bar{\rho}^2 \geq \mu\delta^2$.

Next, we prove the second part of the lemma, i.e., (4.52), pertaining to the S-AW algorithm. Recall that as $\mathcal{C}$ is a polytope, we can write $\mathcal{C} = \operatorname{conv}(\mathcal{A})$ where $\mathcal{A}$ is a finite set of *atoms* in $\mathbb{R}^n$, i.e., $\mathcal{C}$ is a convex hull of $\mathcal{A}$. Note that $\mathcal{A}_t \subseteq \mathcal{A}$ for all t in the S-AW algorithm. Let $\boldsymbol{a}(\mathcal{K}, \boldsymbol{d}) := \arg\max_{\boldsymbol{v}\in\mathcal{K}}\langle\boldsymbol{v}, \boldsymbol{d}\rangle$. Now, define the quantities:

$$\gamma^A(\boldsymbol{\theta}, \boldsymbol{\theta}') := \frac{\langle\nabla f(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle}{\langle\nabla f(\boldsymbol{\theta}), \boldsymbol{v}_f(\boldsymbol{\theta}) - \boldsymbol{s}_f(\boldsymbol{\theta})\rangle} \ , \tag{4.77}$$

where $\boldsymbol{v}_f(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{a}\in\mathcal{A}(\boldsymbol{\theta})}\langle\nabla f(\boldsymbol{\theta}), \boldsymbol{a}\rangle$ and $\boldsymbol{s}_f(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{a}\in\mathcal{A}}\langle\nabla f(\boldsymbol{\theta}), \boldsymbol{a}\rangle$. From (Lacoste-Julien & Jaggi, 2015, Theorem 6), it can be verified that

$$\mu\cdot\delta_{\text{AW}}^2 \leq \inf_{\boldsymbol{\theta}\in\mathcal{C}}\Big(\inf_{\substack{\boldsymbol{\theta}'\in\mathcal{C},\text{s.t.}\\ \langle\nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle<0}}\Big(\frac{2}{\gamma^A(\boldsymbol{\theta}, \boldsymbol{\theta}')^2}\big(f(\boldsymbol{\theta}') - f(\boldsymbol{\theta}) - \langle\nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle\big)\Big)\Big) \ , \tag{4.78}$$

In the above, we have denoted $\mathcal{A}(\boldsymbol{\theta}) := \{\boldsymbol{v} = \boldsymbol{v}_{\mathcal{A}'}(\boldsymbol{\theta}) \ : \ \mathcal{A}' \in \mathcal{A}_{\boldsymbol{\theta}}\}$ where $\boldsymbol{v}_{\mathcal{A}'}(\boldsymbol{\theta}) := \arg\max_{\boldsymbol{a}\in\mathcal{A}'}\langle\nabla f(\boldsymbol{\theta}), \boldsymbol{a}\rangle$. We remark that $\mathcal{A}(\boldsymbol{\theta}_t) \subseteq \mathcal{A}_t$. Note that $\gamma^A(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0$ as long as $\langle\nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle < 0$ is satisfied.

Assume $\boldsymbol{\theta}_t \neq \boldsymbol{\theta}^\star$ and observe that we have $\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t\rangle < 0$, Eq. (4.78) implies

$$\begin{aligned}\frac{\gamma^A(\boldsymbol{\theta}_t, \boldsymbol{\theta}^\star)^2}{2}\mu\delta_{\text{AW}}^2 &\leq f(\boldsymbol{\theta}^\star) - f(\boldsymbol{\theta}_t) - \langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_t\rangle \\ &= -h_t + \gamma^A(\boldsymbol{\theta}_t, \boldsymbol{\theta}^\star)\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{v}_f(\boldsymbol{\theta}_t) - \boldsymbol{s}_f(\boldsymbol{\theta}_t)\rangle \ ,\end{aligned} \tag{4.79}$$

where the equality is found using the definition of $\gamma^A(\boldsymbol{\theta}_t, \boldsymbol{\theta}^\star)$. Define $g_t^{AW} := \max_{\boldsymbol{\theta}\in\mathcal{A}_t}\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}\rangle - \min_{\boldsymbol{\theta}\in\mathcal{C}}\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}\rangle$ and observe that

$$\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{s}_f(\boldsymbol{\theta}_t)\rangle = \min_{\boldsymbol{\theta}\in\mathcal{C}}\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}\rangle \quad \text{and} \quad \langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{v}_f(\boldsymbol{\theta}_t)\rangle \leq \max_{\boldsymbol{\theta}\in\mathcal{A}_t}\langle\nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}\rangle \ . \quad (4.80)$$

Plugging the above into (4.79) yields

$$h_t \leq -\frac{\gamma^A(\boldsymbol{\theta}_t, \boldsymbol{\theta}^\star)^2}{2}\mu\delta_{\text{AW}}^2 + \gamma^A(\boldsymbol{\theta}_t, \boldsymbol{\theta}^\star)g_t^{AW} \leq \frac{(g_t^{AW})^2}{2\delta_{\text{AW}}^2\mu} \ , \quad (4.81)$$

where we have set $\gamma^A(\boldsymbol{\theta}_t, \boldsymbol{\theta}^\star) = g_t^{AW}/(\delta_{\text{AW}}^2\mu)$ similar to the first part of this proof. This concludes the proof for the lower bound on g_t^{AW}. Lastly, it follows from Remark 7, Eq. (20) and Theorem 6 of Lacoste-Julien & Jaggi (2015) that $\mu\delta_{\text{AW}}^2 \leq L\bar{\rho}^2$.

4.6.3 Proof of Theorem 5

Define $\boldsymbol{b}_t := \arg\min_{\boldsymbol{b}\in\mathcal{C}}\langle\nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{b}\rangle$ and $g_t := \max_{\boldsymbol{b}\in\mathcal{C}}\langle\nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{b}\rangle$. We shall split the proof into the two parts — first we prove the bounds in (4.21) and (4.23), then we prove the asymptotic convergence in (4.25).

4.6.3.1 Convergence rates of the FW gap

We prove the bounds for S-FW algorithm. Observe that under H14, the following holds with probability at least $1-\epsilon$ for all $t \geq T_0/2$,

$$\begin{aligned}\langle\nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle &\leq \langle\hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle + \langle\nabla f_t(\boldsymbol{\theta}_t) - \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle \\ &\leq \langle\hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{b}_t - \boldsymbol{\theta}_t\rangle + \langle\nabla f_t(\boldsymbol{\theta}_t) - \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle \\ &= -g_t + \langle\nabla f_t(\boldsymbol{\theta}_t) - \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{b}_t\rangle \leq -g_t + 2\rho\sigma t^{-\alpha} \ ,\end{aligned} \quad (4.82)$$

where the second inequality is due to the optimality of $\boldsymbol{a}_t$ and the last inequality is due to the boundedness of $\mathcal{C}$ and H14. Now, as each of $f_t(\boldsymbol{\theta})$ is L-smooth, we have

$$\begin{aligned}f_t(\boldsymbol{\theta}_{t+1}) &\leq f_t(\boldsymbol{\theta}_t) + \gamma_t\langle\nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{a}_t - \boldsymbol{\theta}_t\rangle + \gamma_t^2 L\bar{\rho}^2/2 \\ &\leq f_t(\boldsymbol{\theta}_t) + \gamma_t \cdot \Big(-g_t + 2\rho\sigma t^{-\alpha} + \gamma_t L\bar{\rho}^2/2\Big) \ .\end{aligned} \quad (4.83)$$

To obtain (4.21), we sum up the both sides of (4.83) from $t = T/2+1$ to $t = T$ to yield,

$$\sum_{t=T/2+1}^{T}\gamma_t g_t \leq \sum_{t=T/2+1}^{T}\gamma_t\big(2\rho\sigma t^{-\alpha} + \frac{L\bar{\rho}^2}{2}\gamma_t\big) + \sum_{t=T/2+1}^{T}\left(f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}_{t+1})\right) . \quad (4.84)$$

As $g_t \geq 0$ for all t, we can lower bound the left hand side by:

$$\sum_{t=T/2+1}^{T} t^{-\eta}g_t \geq \Big(\sum_{t=T/2+1}^{T} t^{-\eta}\Big)\min_{t\in[T/2+1,T]} g_t \geq \frac{T^{1-\eta}}{1-\eta}\left(1 - \left(\frac{2}{3}\right)^{1-\alpha}\right)\min_{t\in[T/2+1,T]} g_t \ ,$$

which holds for all $T \geq 6$. On the other hand, we observe that the first summation in the right hand side of (4.84) can be bounded as

$$\sum_{t=T/2+1}^{T} t^{-\eta}\big(2\rho\sigma t^{-\alpha} + \frac{L\bar{\rho}^2}{2}t^{-\eta}\big) \leq (2\rho\sigma + (L\bar{\rho}^2/2)) \cdot \sum_{t=T/2+1}^{T} t^{-\min\{2\eta,\eta+\alpha\}} \,. \tag{4.85}$$

Meanwhile, the second summation can be bounded by

$$\sum_{t=T/2+1}^{T} (f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}_{t+1})) \leq f_{T/2+1}(\boldsymbol{\theta}_{T/2+1}) - f_T(\boldsymbol{\theta}_{T+1}) + \sum_{t=T/2+1}^{T-1} C_b \cdot t^{-\beta} \,. \tag{4.86}$$

For any $1 > \delta > 0$, the following bound holds

$$\sum_{t=T/2+1}^{T} t^{-\delta} \leq \int_{T/2+1}^{T} t^{-\delta}\, dt \leq \frac{T^{1-\delta}}{1-\delta}\Big(1 - \big(\frac{1}{2}\big)^{1-\delta}\Big) \leq \log 2 \cdot T^{1-\delta} \,. \tag{4.87}$$

For $\delta \geq 1$, we have $\sum_{t=T/2+1}^{T} t^{-\delta} \leq \int_{T/2+1}^{T} t^{-\delta}\, dt \leq \log 2$. Therefore, we can write the upper bound as

$$\sum_{t=\frac{T}{2}+1}^{T} t^{-\delta} \leq T^{\max\{0,1-\delta\}} \cdot \log 2 \,.$$

Notice that the upper bound is decreasing with δ for all $T \geq 2$. Moreover, we have $f_{T/2+1}(\boldsymbol{\theta}_{T/2+1}) - f_T(\boldsymbol{\theta}_{T+1}) \leq 2B < \infty$ from the boundedness of $\mathcal{C}$ and smoothness of f_t. Let $\delta := \min\{2\eta, \eta+\alpha, \beta\}$. The FW gap can be bounded by

$$\begin{aligned} &\min_{t\in[T/2+1,T]} g_t \leq \\ &\left(1 - \left(\frac{2}{3}\right)^{1-\eta}\right)^{-1} \left(2B + \left(C_b + 2\rho\sigma + \frac{L\bar{\rho}^2}{2}\right) \cdot \log 2\right) \cdot T^{-\min\{1-\eta,\delta-\eta\}} \,. \end{aligned} \tag{4.88}$$

Furthermore, we remark that as $\eta > 0.5$, we have $\min\{1-\eta, \delta-\eta\} = \min\{1-\eta, \eta, \alpha, \beta-\eta\} = \min\{1-\eta, \alpha, \beta-\eta\}$.

To prove the improved convergence rate (4.23), we observe that from (4.83), when statement (b-i) in (4.23) is violated, *i.e.,*

$$2\rho\sigma t^{-\eta} + (L\bar{\rho}^2/2)t^{-\alpha} < \max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}\rangle,\ \forall\, t \in [T/2+1, T] \,, \tag{4.89}$$

then $f_t(\boldsymbol{\theta}_{t+1}) < f_t(\boldsymbol{\theta}_t)$ for all $t \in [T/2+1, T]$ and statement (a) holds. Otherwise, when statement (a) is violated, we see that statement (b-i) holds automatically.

We now show the bounds for S-AW algorithm by establishing the analogous inequalities to (4.82) and (4.83). Observe that with probability at least $1-\epsilon$, for all $t \geq T_0/2$,

$$\begin{aligned} \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{d}_t\rangle &\leq \langle \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{d}_t\rangle + \rho\sigma t^{-\alpha} \leq \langle \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t\rangle + \rho\sigma t^{-\alpha} \\ &\leq \langle \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{b}_t - \boldsymbol{\theta}_t\rangle + \rho\sigma t^{-\alpha} \leq -g_t + 2\rho\sigma t^{-\alpha} \,, \end{aligned} \tag{4.90}$$

where we have used the construction that $\langle \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{d}_t\rangle = \min\{\langle \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{a}_t^{\mathsf{FW}} - \boldsymbol{\theta}_t\rangle, \langle \hat{\nabla}_t f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{a}_t^{\mathsf{AW}}\rangle\}$ in the second inequality; the optimality of $\boldsymbol{a}_t^{\mathsf{FW}}$ in the third inequality and the definition

of g_t in the last inequality. Consequently, we have

$$\begin{aligned} f_t(\boldsymbol{\theta}_{t+1}) &\leq f_t(\boldsymbol{\theta}_t) + \hat{\gamma}_t \langle \nabla f_t(\boldsymbol{\theta}_t), \boldsymbol{d}_t \rangle + \frac{1}{2}\hat{\gamma}_t^2 L\bar{\rho}^2 \\ &\leq f_t(\boldsymbol{\theta}_t) + \hat{\gamma}_t\big(-g_t + 2\rho\sigma t^{-\alpha} + \frac{1}{2}\hat{\gamma}_t L\bar{\rho}^2\big) \,. \end{aligned} \tag{4.91}$$

We can obtain a similar bound as (4.84) by summing the both sides from $t = T/2+1$ to $t = T$.

$$\sum_{t=T/2+1}^{T} \hat{\gamma}_t g_t \leq \sum_{t=T/2+1}^{T} \hat{\gamma}_t\big(2\rho\sigma t^{-\alpha} + \frac{L\bar{\rho}^2}{2}\hat{\gamma}_t\big) + \sum_{t=T/2+1}^{T} \left(f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}_{t+1})\right) . \tag{4.92}$$

To lower bound the left hand side, define $\mathcal{T}_{\text{non-drop}}$ be a subset of $[T/2+1, T]$ where a non-drop step is taken. We have

$$\sum_{t=T/2+1}^{T} \hat{\gamma}_t \geq \sum_{t\in\mathcal{T}_{\text{non-drop}}} \gamma_{n_t} \geq \sum_{t=3T/4+1}^{T} \gamma_t \geq \frac{T^{1-\eta}}{1-\eta}\Big(1 - \Big(\frac{4}{5}\Big)^{1-\eta}\Big) \,, \tag{4.93}$$

where the second inequality is due to the fact that $|\mathcal{T}_{\text{non-drop}}| \geq T/4$ (cf. Lemma 2) and the last inequality holds for all $T \geq 20$.

On the other hand, Lemma 2 implies that at least $T/4$ non-drop steps could have taken until round $T/2$, therefore we have $\hat{\gamma}_t \leq \gamma_{T/4}$ for all $t \in [T/2+1, T]$ since if a non-drop step is taken, then the step size will decrease; or if a drop-step step is taken, we have $\hat{\gamma}_t \leq \gamma_{n_{t-1}}$ and $n_{t-1} \geq T/4$. Therefore, we obtain

$$\frac{1}{2}\sum_{t=T/2+1}^{T} \hat{\gamma}_t^2 L\bar{\rho}^2 \leq L\bar{\rho}^2\Big(\frac{T}{4}\Big)^{1-2\eta} \leq L\bar{\rho}^2 \cdot T^{1-2\eta} \,. \tag{4.94}$$

Similarly, we can show that $\sum_{t=T/2+1}^{T} \hat{\gamma}_t 2\rho\sigma t^{-\alpha} \leq 8\rho\sigma \log 2 \cdot T^{1-\alpha-\eta}$. The right hand side of (4.92) can thus be upper bounded by:

$$2B + \big(L\bar{\rho}^2 + 8\rho\sigma + C_b\big) \cdot \log 2 \cdot T^{\max\{0, 1-\beta, 1-2\eta, 1-\alpha-\eta\}} \,. \tag{4.95}$$

Finally, the FW gap for S-AW algorithm can be bounded by

$$\begin{aligned} &\min_{t\in[T/2+1,T]} g_t \leq \\ &\left(1 - \left(\frac{4}{5}\right)^{1-\eta}\right)^{-1} \big(2B + \big(L\bar{\rho}^2 + 8\rho\sigma + C_b\big) \cdot \log 2\big) \cdot T^{-\min\{1-\eta, \beta-\eta, \alpha\}} \,. \end{aligned} \tag{4.96}$$

Moreover, we can prove (4.24) by using the same argument for S-FW on (4.92).

4.6.3.2 Asymptotic Convergence

We now show the asymptotic convergence to a stationary point for the S-FW/S-AW algorithms. Define the following set of stationary points to (4.1):

$$\mathcal{C}^\star := \text{co}(\{\bar{\boldsymbol{\theta}} \in \mathcal{C} : \max_{\boldsymbol{\theta}\in\mathcal{C}} \langle \nabla f(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle = 0\}) \,, \tag{4.97}$$

and $\mathrm{co}(\cdot)$ denotes the closure of the set. We shall invoke the following Nurminskii's sufficient condition:

Theorem 15. *(Nurminskii, 1972, Theorem 1) For an arbitrary convergent subsequence $\{\boldsymbol{\theta}_{s_k}\}_{k\geq 1}$ in $\mathcal{C}$ with the limit point $\underline{\boldsymbol{\theta}}$, if the following conditions hold:*

1. *if $\underline{\boldsymbol{\theta}} \in \mathcal{C}^\star$, then $\lim_{k\to\infty} \|\boldsymbol{\theta}_{s_k+1} - \boldsymbol{\theta}_{s_k}\| = 0$.*
2. *if $\underline{\boldsymbol{\theta}} \notin \mathcal{C}^\star$, then there exists $\epsilon_0 > 0$ such that for all $0 < \epsilon \leq \epsilon_0$, the integer quantity τ_k is finite with*
$$\tau_k := \min_{s>s_k} \; s \;\; \text{s.t.} \;\; \|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s_k}\| > \epsilon \; . \tag{4.98}$$
3. *taking the same τ_k defined above, there exists a continuous function $W(\boldsymbol{\theta})$ that takes a finite number of values in $\mathcal{C}^\star$ such that*
$$\overline{\lim_{k\to\infty}} W(\boldsymbol{\theta}_{\tau_k}) < \lim_{k\to\infty} W(\boldsymbol{\theta}_{s_k}) \; . \tag{4.99}$$

Then the sequence $\{W(\boldsymbol{\theta}_k)\}_{k\geq 1}$ converges and the limit points of the sequence $\{\boldsymbol{\theta}_t\}_{t\geq 1}$ belongs to the set $\mathcal{C}^\star$.

Our plan is to apply the theorem above to prove (4.25). We first observe that as $\mathcal{C}$ is closed and bounded, by Bolzano-Weierstrass theorem there exists a convergent subsequence $\{\boldsymbol{\theta}_{s_k}\}_{k\geq 1}$ of the sequence of iterates generated by the S-FW/S-AW algorithms. Moreover, for S-FW algorithm, Condition 1) can be easily verified since

$$\|\boldsymbol{\theta}_{s_k+1} - \boldsymbol{\theta}_{s_k}\| \leq \gamma_{s_k} \|\boldsymbol{a}_t - \boldsymbol{\theta}_{s_k}\| \leq \gamma_{s_k} \bar{\rho} \; , \tag{4.100}$$

and $\gamma_{s_k} \to 0$ as $k \to \infty$. This also holds for the S-AW algorithm by replacing the above γ_{s_k} by $\hat{\gamma}_{s_k}$.

Now, let $\underline{\boldsymbol{\theta}}$ be the limit of the subsequence $\{\boldsymbol{\theta}_{s_k}\}_{k\geq 1}$ and $\underline{\boldsymbol{\theta}} \notin \mathcal{C}$. We shall verify condition 2) in Theorem 15 by contradiction. In particular, we assume that the following holds for all $0 < \epsilon \leq \epsilon_0$:

$$\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s_k}\| \leq \epsilon, \; \forall \; s > s_k \; . \tag{4.101}$$

For some sufficiently large k and $s > s_k$, since $\boldsymbol{\theta}_{s_k}$ converges to $\underline{\boldsymbol{\theta}}$ as $k \to \infty$, we have $\boldsymbol{\theta}_s \in \mathcal{B}_{2\epsilon}(\underline{\boldsymbol{\theta}})$, i.e., the ball of radius 2ϵ centered at $\underline{\boldsymbol{\theta}}$. Furthermore, as $\underline{\boldsymbol{\theta}} \notin \mathcal{C}^\star$ and $\mathcal{C}^\star$ is closed, the following holds,

$$\langle \nabla f(\boldsymbol{\theta}_s), \boldsymbol{\theta} - \boldsymbol{\theta}_s \rangle \leq -\delta < 0, \; \forall \; \boldsymbol{\theta} \in \mathcal{C}, \; \forall \; s > s_k \; , \tag{4.102}$$

for some $\delta > 0$. In particular, we have $\langle \nabla f(\boldsymbol{\theta}_s), \boldsymbol{b}_s - \boldsymbol{\theta}_s \rangle \leq -\delta$. From (4.83), H14 and H15, it holds true for all $t \geq 1$ that:

$$f(\boldsymbol{\theta}_{t+1}) - f(\boldsymbol{\theta}_t) \leq \gamma_t \cdot \langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{b}_t - \boldsymbol{\theta}_t \rangle + \gamma_t 2\rho\sigma \cdot t^{-\alpha} + \frac{1}{2}\gamma_t^2 L \bar{\rho}^2 \; . \tag{4.103}$$

To arrive at a contradiction, we let $s > s_k$ and sum up the both side of the above from $t = s_k$ to $t = s$. Consider the following chain of inequality:

$$\begin{aligned} f(\boldsymbol{\theta}_s) - f(\boldsymbol{\theta}_{s_k}) &\leq \sum_{\ell=s_k}^{s} \gamma_\ell \cdot (\langle \nabla f(\boldsymbol{\theta}_\ell), \boldsymbol{a}_\ell - \boldsymbol{\theta}_\ell \rangle + 2\rho\sigma \cdot t^{-\alpha} + (L\bar{\rho}^2/2) \cdot t^{-\eta}) \\ &\leq -\delta \sum_{\ell=s_k}^{s} \gamma_\ell + \sum_{\ell=s_k}^{s} \ell^{-\eta} (2\rho\sigma \cdot \ell^{-\alpha} + (L\bar{\rho}^2/2) \cdot \ell^{-\eta}) \; , \end{aligned} \tag{4.104}$$

where the second inequality is due to (4.102). Letting $s \to \infty$ and observe that $\sum_{\ell=s_k}^{\ell} \gamma_\ell \to +\infty$ implies

$$\lim_{s\to\infty} f(\bar{\boldsymbol{\theta}}^s) - f(\bar{\boldsymbol{\theta}}^{s_k}) < -\infty \,, \tag{4.105}$$

since $\lim_{s\to\infty} \sum_{\ell=s_k}^{s} \ell^{-\eta}(2\rho\sigma \cdot \ell^{-\alpha} + (L\bar{\rho}^2/2) \cdot \ell^{-\eta}) < \infty$, which is due to $\eta + \alpha > 1$ and $2\eta > 1$. This leads to a contradiction since $f(\boldsymbol{\theta})$ is bounded over $\mathcal{C}$. We conclude that condition 2) holds for the S-FW algorithm. Due to Lemma 2, the same arguments can be repeated for the S-AW algorithm.

The remaining task is to verify condition 3) in Theorem 15. Here we shall take $W(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$. By the definition of τ_k, we have $\boldsymbol{\theta}_s \in \mathcal{B}_\epsilon(\boldsymbol{\theta}_{s_k})$ for all $s_k \le s \le \tau_k - 1$. Again for some sufficiently large k, we have $\boldsymbol{\theta}_s \in \mathcal{B}_\epsilon(\boldsymbol{\theta}_{s_k}) \subseteq \mathcal{B}_{2\epsilon}(\underline{\boldsymbol{\theta}})$ and the inequality (4.104) holds for $s = \tau_k - 1$. This gives:

$$f(\boldsymbol{\theta}_{\tau_k}) - f(\boldsymbol{\theta}_{s_k}) \le \sum_{\ell=s_k}^{\tau_k-1} \gamma_\ell \cdot (-\delta + 2\rho\sigma \cdot \ell^{-\alpha} + (L\bar{\rho}^2/2) \cdot \ell^{-\eta}) \,. \tag{4.106}$$

On the other hand, we have $\boldsymbol{\theta}_{\tau_k} \notin \mathcal{B}_\epsilon(\boldsymbol{\theta}_{s_k})$ and thus

$$\epsilon < \|\boldsymbol{\theta}_{\tau_k} - \boldsymbol{\theta}_{s_k}\| \le \sum_{\ell=s_k}^{\tau_k-1} \gamma_\ell \|\hat{\boldsymbol{a}}_\ell - \boldsymbol{\theta}_\ell\| \le \bar{\rho} \sum_{\ell=s_k}^{\tau_k-1} \gamma_\ell \,. \tag{4.107}$$

The above implies that $\sum_{\ell=s_k}^{\tau_k-1} \gamma_\ell > \epsilon/\bar{\rho} > 0$. Considering (4.106) again, observe that the latter two terms decay to zero, for some sufficiently large k, therefore we have $-\delta + \mathcal{O}(\ell^{-\min\{\eta,\alpha\}}) \le -\delta' < 0$ if $\ell \ge s_k$ for some $\delta' > 0$. Finally, (4.106) leads to

$$f(\boldsymbol{\theta}_{\tau_k}) - f(\boldsymbol{\theta}_{s_k}) \le -\delta' \sum_{\ell=s_k}^{\tau_k-1} \gamma_\ell < -\frac{\delta'\epsilon}{\bar{\rho}} < 0 \,. \tag{4.108}$$

Taking the limit $k \to \infty$ on the both sides lead to (4.99) and completes the proof. Again, the same arguments can be repeated for the S-AW algorithm due to Lemma 2 that controls the magnitude of $\hat{\gamma}_t$.

4.6.4 Proof of Proposition 6

The following proof relies on a modified version of (Shalev-Shwartz et al., 2009, Theorem 5)[3]. Let us define

$$\boldsymbol{\epsilon}_t(\boldsymbol{\theta}) = \nabla F_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}) = \frac{1}{t} \sum_{s=1}^{t} \Big(\nabla f(\boldsymbol{\theta}; \omega_s) - \mathbb{E}_{\omega\sim\mathcal{D}}[\nabla f(\boldsymbol{\theta}; \omega)] \Big) \,. \tag{4.109}$$

From Gaugry (2005), there exists an Euclidean $(\epsilon/(L\bar{\rho}))$-net $\mathcal{N}(\epsilon)$ with cardinality less than

$$|\mathcal{N}(\epsilon)| = \mathcal{O}\left(n^2 \log(n) \left(\frac{L\bar{\rho}}{\epsilon} \right)^n \right)$$

[3]Note that (Shalev-Shwartz et al., 2009, Theorem 5) has implicitly assumed $\mathcal{D}$ to have a bounded support, which is more restrictive than our setting here.

Therefore, for any $\boldsymbol{\theta} \in \mathcal{C}$ there is a point $p_\theta \in \mathcal{N}$ such that:

$$\begin{aligned}\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_\infty &\leq \|\boldsymbol{\epsilon}_t(p_{\boldsymbol{\theta}})\|_\infty + \|\boldsymbol{\epsilon}_t(p_{\boldsymbol{\theta}}) - \boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_\infty \\ &\leq \|\boldsymbol{\epsilon}_t(p_{\boldsymbol{\theta}})\|_\infty + \|\nabla F_t(\boldsymbol{\theta}) - \nabla F_t(p_{\boldsymbol{\theta}})\|_\infty + \|\nabla F(\boldsymbol{\theta}) - \nabla F(p_{\boldsymbol{\theta}})\|_\infty \\ &\leq \|\boldsymbol{\epsilon}_t(p_{\boldsymbol{\theta}})\|_\infty + 2\epsilon\end{aligned} \tag{4.110}$$

where we used the Lipschitz assumption for the second inequality. Controlling each point $\mathcal{N}$ using the sub-Gaussian assumption and applying the union bound yields:

$$\begin{aligned}\mathbb{P}(|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})| > s) \leq \mathbb{P}(|\boldsymbol{\epsilon}_t(p_{\boldsymbol{\theta}})| > s - 2\epsilon) &\leq 2|\mathcal{N}(\epsilon)| \exp\left(-\frac{t(s-2\epsilon)^2}{2\sigma_D^2}\right) \\ &\leq \mathcal{O}\left(n^2 \log(n)(\frac{L\bar{\rho}}{\epsilon})^n \exp\left(-\frac{t(s-2\epsilon)^2}{2\sigma_D^2}\right)\right)\end{aligned}$$

If $n \geq 4$, setting $s = 3\epsilon$ shows that with probability at least $1 - \delta$

$$\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_\infty = \mathcal{O}\left(\max(L\bar{\rho},\ \sigma_D)\sqrt{\frac{n\log(t)\log(n/\delta)}{t}}\right) \tag{4.111}$$

Applying the union bound over t then yields the result.

4.6.5 Proof of Proposition 10

Notice that the gradient vector is given by:

$$\nabla f(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{Y})] = \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{Y}]. \tag{4.112}$$

We can bound the gradient estimation error as:

$$\|\nabla F_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})\|_\infty \leq \Big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{A}_s^\top \boldsymbol{w}_s\Big\|_\infty + \Big\|\frac{1}{t}\sum_{s=1}^t \big(\boldsymbol{A}_s^\top\boldsymbol{A}_s - \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]\big)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\Big\|_\infty \tag{4.113}$$

To bound the second term in (4.113), we define $\boldsymbol{Z}_s := \boldsymbol{A}_s^\top\boldsymbol{A}_s - \mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]$. Observe that

$$\Big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{Z}_s(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\Big\|_\infty = \max_{i\in[n]}\Big|\frac{1}{t}\sum_{s=1}^t \boldsymbol{z}_{s,i}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\Big|, \tag{4.114}$$

where $\boldsymbol{z}_{s,i}$ denotes the ith row vector in $\boldsymbol{Z}_s$. Furthermore, by the Holder's inequality,

$$\Big|\frac{1}{t}\sum_{s=1}^t \boldsymbol{z}_{s,i}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\Big| \leq \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_1 \Big\|\frac{1}{t}\sum_{s=1}^t \boldsymbol{z}_{s,i}\Big\|_\infty, \tag{4.115}$$

Now that $\boldsymbol{z}_{s,i}$ is a zero-mean, independent random vector with elements bounded in $[-B_1, B_1]$, applying the union bound and the Hoefding's inequality gives:

$$\mathbb{P}\big(\big\|(1/t)\textstyle\sum_{s=1}^t \boldsymbol{z}_{s,i}\big\|_\infty \geq x,\ \forall\, i\big) \leq 2n^2 \mathrm{e}^{-\frac{x^2 t}{2B_1^2}}. \tag{4.116}$$

Setting $x = B_1\sqrt{2(\log(2n^2t^2) - \log\epsilon)/t}$ gives ϵ/t^2 on the right hand side. With probability at least $1 - \epsilon/t^2$, we have

$$\Big\|\frac{1}{t}\sum_{s=1}^{t}\boldsymbol{Z}_s\boldsymbol{\theta}\Big\|_\infty \leq cB_1\sqrt{2(\log(2n^2t^2) - \log\Delta)/t}, \tag{4.117}$$

To bound the first term in (4.113), we find that the ith element of the vector $\boldsymbol{A}_s^\top\boldsymbol{w}_s$ is zero-mean. Furthermore, it can be verified that

$$\mathbb{E}\big[\mathrm{e}^{\left(\lambda\sum_{j=1}^{m}A_{s,i,j}w_{s,j}\right)}\big] \leq \mathrm{e}^{\lambda^2\cdot m\sigma_w^2B_2/2}, \tag{4.118}$$

for all $\lambda \in \mathbb{R}$, where $A_{s,i,j}$ is the (i,j)th element of $\boldsymbol{A}_s$ and $w_{s,j}$ is the jth element of $\boldsymbol{w}_s$. In other words, the ith element of $\boldsymbol{A}_s^\top\boldsymbol{w}_s$ is sub-Gaussian with parameter $m\cdot\sigma_w^2B_2$. It follows by the Hoefding's inequality that

$$\mathbb{P}\big(\big\|\frac{1}{t}\sum_{s=1}^{t}\boldsymbol{A}_s^\top\boldsymbol{w}_s\big\|_\infty \geq x\big) \leq 2n\mathrm{e}^{-\frac{x^2t}{2mB_2\sigma_w^2}}. \tag{4.119}$$

Setting $x = \sigma_w\sqrt{2mB_2(\log(2n^2t^2) - \log\epsilon)/t}$ yields $\epsilon/(nt^2)$ on the right hand side. Combining (4.116), (4.119) and using a union bound argument (for all $t \geq 1$) yields the desired result.

4.6.6 Proof of Proposition 11

We first state and prove the following technical result that is necessary for our proof.

Proposition 16. *Consider a finite sequence of independent random matrices* $(Z_s)_{1\leq s\leq t} \in \mathbb{R}^{m_1\times m_2}$ *satisfying* $\mathbb{E}[Z_i] = 0$. *For some* $U > 0$, *assume*

$$\inf\{\lambda > 0 : \mathbb{E}[\exp(\|Z_i\|_{\sigma,\infty}/\lambda)] \leq \mathrm{e}\} \leq U \quad \forall i \in [n], \tag{4.120}$$

and there exists σ_Z *s.t.*

$$\sigma_Z^2 \geq \max\left\{\left\|\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}[Z_sZ_s^\top]\right\|_{\sigma,\infty}, \left\|\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}[Z_s^\top Z_s]\right\|_{\sigma,\infty}\right\}. \tag{4.121}$$

Then for any $\nu > 0$, *with probability at least* $1 - \mathrm{e}^{-\nu}$

$$\left\|\frac{1}{t}\sum_{i=1}^{t}Z_i\right\|_{\sigma,\infty} \leq c_U\max\left\{\sigma_Z\sqrt{\frac{\nu+\log(d)}{t}}, U\log(\frac{U}{\sigma_Z})\frac{\nu+\log(d)}{t}\right\}, \tag{4.122}$$

with c_U *an increasing constant with* U.

Proof. This result is proved in Theorem 4 in Koltchinskii (2013) for symmetric matrices. Here we state a slightly different result because σ_Z^2 is an upper bound of the variance and not the variance itself. However, it does not the alter the proof and the result stays valid. This concentration is extended to rectangular matrices by dilation, see Proposition 11 in Klopp (2014) for details. □

Define $\boldsymbol{\epsilon}_t(\boldsymbol{\theta}) = \nabla F_t(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta})$. To prove Proposition 11, observe that for a fixed $\boldsymbol{\theta}$ and using the triangle inequality

$$\begin{aligned}\|\boldsymbol{\epsilon}_t(\boldsymbol{\theta})\|_{\sigma,\infty} \leq \|\frac{1}{t}\sum_{s=1}^{t} Y_s e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[Y_s e_{k_s} e_{l_s}^{'\top}]\|_{\sigma,\infty} + \\ \|\frac{1}{t}\sum_{s=1}^{t} \mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[\mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top}]\|_{\sigma,\infty} \,.\end{aligned}$$

Define $Z_s := Y_s e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[Y_s e_{k_s} e_{l_s}^{'\top}]$, then

$$\begin{aligned}&\|\mathbb{E}[Z_s Z_s^\top]\|_{\sigma,\infty} \leq \|\mathbb{E}[Y_s^2 e_{k_s} e_{l_s}^{'\top} e_{l_s}^{'} e_{k_s}^\top]\|_{\sigma,\infty} \,, \\ &= \|\frac{1}{m_1 m_2} \operatorname{diag}\left(\left(\sum_{l=1}^{m_2} \mathbb{E}[Y_s^2|k,l]\right)_{k=1}^{m_1}\right)\|_{\sigma,\infty} \,, \\ &= \frac{1}{m_1 m_2} \max_{k\in[m_1]} \left(\sum_{l=1}^{m_2} \mathrm{A}''(\bar{X}_{k,l}) + (\mathrm{A}'(\bar{X}_{k,l}))^2\right) \,, \\ &\leq \frac{\bar{\sigma}^2}{m_1 \wedge m_2} + \frac{\kappa^2}{m_1 m_2} \leq \frac{\bar{\sigma}^2 + \kappa^2}{m_1 \wedge m_2} \,,\end{aligned}$$

where we used the fact that the distribution belongs to the exponential family for the second equality. Similarly one shows that $\|\mathbb{E}[Z_s^\top Z_s]\|_{\sigma,\infty}$ satisfies the same upper bound. Hence by Proposition 16 and A1, with probability at least $1 - \mathrm{e}^{-\nu}$, it holds

$$\|\frac{1}{t}\sum_{s=1}^{t} Z_s\|_{\sigma,\infty} \leq c_\lambda \sqrt{\frac{(\bar{\sigma}^2 + \kappa^2)(\nu + \log(d))}{t(m_1 \wedge m_2)}} \,, \tag{4.123}$$

for t larger than the threshold given in the proposition statement. For the second term, define $P_t := 1/t \sum_{s=1}^{t} e_{k_s} e_{l_s}^{'\top} - (m_1 m_2)^{-1} \mathbf{1}\mathbf{1}^\top$, we get

$$\begin{aligned}&\|\frac{1}{t}\sum_{s=1}^{t} \mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top} - \mathbb{E}[\mathrm{A}'(\boldsymbol{\theta}_{k_s,l_s}) e_{k_s} e_{l_s}^{'\top}]\|_{\sigma,\infty} \\ &\qquad = \|P_t \odot (\mathrm{A}'(\boldsymbol{\theta}_{k,l}))_{k,l}\|_{\sigma,\infty} \leq \kappa \|P_t\|_{\sigma,\infty} \,,\end{aligned} \tag{4.124}$$

where $\odot$ denotes the Hardamard product and we have used Theorem 5.5.3 in Horn & Johnson (1994) for the last inequality. Define $Z_s' := e_{k_s} e_{l_s}^{'\top} - (m_1 m_2)^{-1} \mathbf{1}\mathbf{1}^\top$. Since by definition, $\lambda \geq 1$, one can again apply Proposition 16 for $U = \lambda$ and get with probability at least $1 - \mathrm{e}^{-\nu}$,

$$\|P_t\|_{\sigma,\infty} \leq c_\lambda \sqrt{\frac{\nu + \log(d)}{t(m_1 \wedge m_2)}} \,. \tag{4.125}$$

Hence, by a union bound argument we find that with probability at least $1 - 2\mathrm{e}^{-\nu}$

$$\|\boldsymbol{\epsilon}_t\|_{\sigma,\infty} \leq c_\lambda \left(2\kappa + \bar{\sigma}\right) \sqrt{\frac{\nu + \log(d)}{t(m_1 \wedge m_2)}} \,. \tag{4.126}$$

Taking $\nu = \log(1 + 2t^2/\epsilon)$ and applying a union bound argument yields the result.

4.6.7 Fast convergence of S-AW without strong convexity

The proof is based on a generalization of Lemma 13, and the following result is borrowed from Theorem 11 in Lacoste-Julien & Jaggi (2015). In particular, we consider the conditions when *(i)* $\mathcal{C}$ is a polytope and *(ii)* the loss function can be written as:

$$f(\boldsymbol{\theta}) = g(\boldsymbol{A}\boldsymbol{\theta}) + \langle \boldsymbol{b}, \boldsymbol{\theta} \rangle \ . \tag{4.127}$$

where g is μ_g-strongly convex. Note that for a general matrix $\boldsymbol{A}$, $f(\boldsymbol{\theta})$ may not be strongly convex.

Define $\boldsymbol{C}$ to be the matrix with rows containing the linear inequalities defining $\mathcal{C}$. Let c_h be the Hoffman constant Lacoste-Julien & Jaggi (2015) for the matrix $[\boldsymbol{A}; \boldsymbol{b}^\top; \boldsymbol{C}]$, $G = \max_{\boldsymbol{\theta} \in \mathcal{C}} \|\nabla g(\boldsymbol{A}\boldsymbol{\theta})\|$ be the maximal norm of gradient of g over $\boldsymbol{A}\mathcal{C}$, $\rho_{\boldsymbol{A}}$ be the diameter of $\boldsymbol{A}\mathcal{C}$ and we define the generalized strong convexity constant:

$$\tilde{\mu} := \frac{1}{2c_h^2(\|\boldsymbol{b}\|M + 3G\rho_{\boldsymbol{A}} + (2/\mu_g)(G^2+1)} \ . \tag{4.128}$$

Under H13 and assuming that $h_t > 0$ holds, applying the inequality (43) from Lacoste-Julien & Jaggi (2015) yields

$$\bar{g}_t^{\mathsf{AW}} \geq \delta_{AW} \sqrt{2\tilde{\mu} \cdot h_t} \ , \tag{4.129}$$

where $\bar{g}_t^{\mathsf{AW}}$ is defined in (4.62). The above result is analogous to that of Lemma 13. Subsequently, the fast convergence results in Theorem 4 can be obtained by repeating the proof in Section 4.6.2.2 with (4.129).

Bibliography

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, 2010.

Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *ICML*, 2016.

Y. F. Atchade, G. Fort, and E. Moulines. On stochastic proximal gradient algorithms. *arXiv preprint arXiv:1402.2365*, 2014.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009a.

A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery. *Convex optimization in signal processing and communications*, pages 42–88, 2009b.

Robert M. Bell and Yehuda Koren. Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl.*, 9(2):75–79, 2007.

R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46(0):109 – 132, 2013. ISSN 0950-7051.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *in COMPSTAT*, 2010.

S. Burer and D.C. R Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103:427–444, 2005.

J-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

T. T. Cai and W-X. Zhou. Matrix completion via max-norm constrained optimization. *CoRR*, abs/1303.0341, 2013a.

T. T. Cai and W-X. Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.*, 14:3619–3647, 2013b.

E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080, 2010.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Sys. and Tech.*, 2:27:1–27:27, 2011.

P. L. Combettes and J. C. Pesquet. Proximal Splitting Methods in Signal Processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *CoRR*, abs/1209.3672, 2012.

Mark A. Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-Bit matrix completion. *Inf. Inference*, 3, 2014.

Marco Duarte, Mark Davenport, Dharmpal Takhar, Jason Laska, Ting Sun, Kevin Kelly, and Richard Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, Mar 2008.

M. Dudík, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *AISTATS*, 2012.

Yu. M. Ermol'ev and P. I. Verchenko. A linearization method in limiting extremal problems. *Cybernetics*, 12(2):240–245, 1976. ISSN 1573-8337.

S. Ertekin, L. Bottou, and C. Lee Giles. Nonconvex online support vector machines. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(2), Feb 2011.

M. Fazel. *Matrix rank minimization with applications*. PhD thesis, 2002.

R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. In *NIPS*, pages 2133–2141, 2011.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logis. Quart.*, 1956.

S. Gaïffas and G. Lecué. Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Trans. Inform. Theory*, 57:6942–6957, 2011.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. *ICML*, 2015a.

D. Garber and E. Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *CoRR*, abs/1301.4666, August 2015b.

Jean-Louis Verger Gaugry. Covering a ball with smaller equal balls in n. *Discrete and Computational Geometry*, 33:143–155, 2005.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *COLT*, 2015.

S. Ghosh and H. Lam. Computing worst-case input models in stochastic simulation. *CoRR*, abs/1507.05609, July 2015.

G. H. Golub and C. F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

G. H. Golub and C. F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. ISBN 978-1-4214-0794-4; 1-4214-0794-9; 978-1-4214-0859-0.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. *ICML*, 2014.

Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for machine learning. In *NIPS workshop*, 2012.

M. Hardt. On the provable convergence of alternating minimization for matrix completion. *CoRR*, 2013.

F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM TiiS*, Jan 2015.

E. Hazan and S. Kale. Projection-free online learning. *ICML*, 2012.

E. Hazan and H. Luo. Variance-reduced and projection-free stochastic optimization. In *ICML*, 2016.

R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

C.-J. Hsieh and P. A. Olsen. Nuclear norm minimization via active subspace selection. In *ICML*, 2014.

C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS*, pages 781–789. 2009.

J. Hui, L. Chaoqiang, S. Zuowei, and X. Yuhong. Robust video denoising using low rank matrix completion. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0: 1791–1798, 2010.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. *ICML*, 2013.

P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, 2013.

L. Ji, P. Musialski, P. Wonka, and Y. Jieping. Tensor Completion for Estimating Missing Values in Visual Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (1):208–220, 2013. ISSN 0162-8828.

Bo Jiang, Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. Structured nonconvex and nonsmooth optimization: Algorithms and iteration complexity analysis. *CoRR*, May 2016.

A. B. Juditsky and A. S. Nemirovski. *First-Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods*. 2012a.

A. B. Juditsky and A. S. Nemirovski. *First-Order Methods for Nonsmooth Convex Large-Scale Optimization, II: Utilizing Problem's Structure*. 2012b.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

Raghunandan Hulikal Keshavan. *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University, 2012.

O. Klopp. Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, 5:1161–1183, 2011.

O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 2 (1):282–303, 02 2014.

O. Klopp, J. Lafond, E. Moulines, and J. Salmon. Adaptive Multinomial Matrix Completion. August 2014.

V. Koltchinskii. *A remark on low rank matrix recovery and noncommutative Bernstein type inequalities*, volume Volume 9 of *Collections*, pages 213–226. Institute of Mathematical Statistics, 2013.

V. Koltchinskii, A. B. Tsybakov, and K. Lounici. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

S. Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *CoRR*, July 2016.

S. Lacoste-Julien and M. Jaggi. An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *NIPS*, 2013.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *NIPS*. 2015.

J. Lafond. Low Rank Matrix Completion with Exponential Family Noise. 2015.

J. Lafond, O. Klopp, E. Moulines, and J. Salmon. Probabilistic low-rank matrix completion on finite alphabets. In *NIPS*. 2014.

J. Lafond, H.T. Wai, and E. Moulines. On the stochastic frank-wolfe algorithms for convex and non-convex optimizations. *ArXiv*, 2016.

G. Lan and Y. Zhou. Conditional gradient sliding for convex optimization. *Tech. Report*, 2014.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23. Springer-Verlag, Berlin, 1991.

P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28, 2000.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322, 2010.

S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13, 2012.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 2009.

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

EA Nurminskii. Convergence conditions for nonlinear programming algorithms. *Cybernetics and Systems Analysis*, pages 959–962, 1972.

N. Parikh, S. Boyd, E. Chu, B. Peleato, and J. Eckstein. Proximal algorithms. *Foundations and Trends in Machine Learning*, 1(3):1–108, 2013.

M. Raginsky and A. Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Trans. Inf. Theory*, 57(10):7036–7056, October 2011.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, December 2011.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

L. Rosasco, S. Villa, and Bang Cong Vu. Convergence of Stochastic Proximal Gradient Algorithm. *CoRR*, abs/1403.5074v3, 2014.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. *COLT*, 2009.

S. Shalev-Shwartz, Ohad Shamir, and Karthik Sidharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM J. Comput.*, 40(6):1623–1646, 2011.

N. Srebro. *Learning with Matrix Factorization*. PhD thesis, 2004.

N. Srebro and R. R. Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. 2010.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

P. Wolfe. Convergence theory in nonlinear programming. *Integer and Nonlinear Program.*, 1970.

L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

H. Xu, W. Jiasong, W. Lu, C. Yang, Lotfi Senhadji, and Huazhong Shu. Linear Total Variation Approximate Regularized Nuclear Norm Optimization for Matrix Completion. *Abstr. Appl. Anal.*, pages Art. ID 765782, 8, 2014. ISSN 1085-3375.

Y. Yang, J. Ma, and S. Osher. Seismic data reconstruction via matrix completion. *Inverse Probl. Imaging*, 7(4):1379–1392, 2013. ISSN 1930-8337.

Yaoliang Yu, Xinhua Zhang, and Dale Schuurmans. Generalized conditional gradient for sparse estimation. *CoRR*, Oct 2014.

C. H. Zhang and T. Zhang. A General Framework of Dual Certificate Analysis for Structured Sparse Recovery Problems. *arXiv.org*, January 2012.

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

Titre : Complétion de Matrice de Faible Rang: Aspects Statistiques et Computationnels

Mots clefs : Statistique en Grande Dimension,Complétion de Matrice,Apprentissage à Grande échelle

Résumé : Dans cette thèse nous nous intéressons aux méthodes de complétion de matrices de faible rang et étudions certains problèmes reliés. Un premier ensemble de résultats visent à étendre les garanties statistiques existantes pour les modèles de complétion avec bruit additif sous-gaussiens à des distributions plus générales. Nous considérons en particulier les distributions multinationales et les distributions appartenant à la famille exponentielle. Pour ces dernières, nous prouvons l'optimalité (au sens minimax) à un facteur logarithmique près des estimateurs à pénalité norme trace. Un second ensemble de résultats concernent l'algorithme du gradient conditionnel qui est notamment utilisé pour calculer les estimateurs précédents. Nous considérons en particulier deux algorithmes de type gradient conditionnel dans le cadre de l'optimisation stochastique. Nous donnons les conditions sous lesquelles ces algorithmes atteignent les performance des algorithmes de type gradient projeté.

Title : Low Rank Matrix Completion: Statistical and Computational Aspects

Keywords : High Dimension Statistics,Matrix Completion,Large Scale Optimization

Abstract : This thesis deals with the low rank matrix completion methods and focuses on some related problems, of both statistical and algorithmic nature. The first part of this work extends the existing statistical guarantees obtained for sub-Gaussian additive noise models, to more general distributions. In particular, we provide upper bounds on the prediction error of trace norm penalized estimator with high probability for multinomial distributions and for distributions belonging to the exponential family. For the latter, we prove that the trace norm penalized estimators are minimax optimal up to a logarithmic factor by giving a lower bound. The second part of this work focuses on the conditional gradient algorithm, which is used in particular to compute previous estimators. We consider the stochastic optimization framework and gives the convergence rate of two variants of the conditional gradient algorithm. We gives the conditions under which these algorithms match the performance of projected gradient algorithms.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France