

# *Tous-sauf-un*

Jeudi 12 novembre 2020

## 1 Prologue : le *tous-sauf-un* en apprentissage

Cette séance vise à présenter une méthode permettant d’analyser théoriquement certains algorithmes non-convexes de reconstruction de matrices de bas rang, plus sophistiquée que celles vues à la séance précédente. En anglais, cette méthode s’appelle *leave-one-out*. En français, on utilise généralement le nom anglais mais l’appellation *tous-sauf-un* se rencontre parfois ; c’est celle-ci que j’utiliserai.

À ma connaissance, cette méthode a été employée pour la première fois en optimisation non-convexe dans l’article [Zhong and Boumal, 2018]. Elle avait auparavant été appliquée à des problèmes statistiques consistant à estimer, à partir d’échantillons d’une variable aléatoire, des paramètres de la loi de cette variable, afin de comprendre certaines propriétés fines des estimateurs proposés [El Karoui, Bean, Bickel, Lim, and Yu, 2013]. Elle repose sur une idée assez simple, employée depuis longtemps dans le domaine de l’apprentissage. Décrivons-la brièvement.

Dans un problème d’apprentissage classique, on souhaite prédire, pour des données  $x \in \mathbb{R}^n$  distribuées selon une loi de probabilité inconnue  $\mathbb{P}$ , une certaine propriété, modélisée par un nombre réel. La propriété est une fonction déterministe des données : elle s’écrit sous la forme  $f_*(x)$ , pour une certaine fonction  $f_* : \mathbb{R}^n \rightarrow \mathbb{R}$ . Pour apprendre à prédire cette propriété, on dispose de plusieurs données  $x_1, \dots, x_m$  (choisies indépendamment selon la loi  $\mathbb{P}$ ) ainsi que des propriétés associées  $f_*(x_1), \dots, f_*(x_m)$ . On envoie ces données et propriétés à un algorithme Alg, qui renvoie une fonction de prédiction  $f_{\text{Alg}}$ .

On mesure fréquemment les performances de l’algorithme par son *erreur de généralisation*

$$\text{Err}(f_{\text{Alg}}) = \mathbb{E}_{x \sim \mathbb{P}}(\ell(f_{\text{Alg}}(x), f_*(x))),$$

où  $\ell : \mathbb{R}^2 \rightarrow \mathbb{R}^+$  désigne une fonction de coût adaptée au problème considéré. Malheureusement, calculer exactement l’erreur de généralisation nécessiterait de connaître  $\mathbb{P}$  et  $f_*$ . Comment l’estimer à partir seulement des données  $x_1, \dots, x_m$  et des propriétés associées ?

Une possibilité est de calculer, pour tout  $i$ , la fonction  $f_{\text{Alg},-i}$  renvoyée par l'algorithme si on lui fournit seulement les données  $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_m$  et les propriétés  $f_*(x_1), \dots, f_*(x_{i-1}), f_*(x_{i+1}), \dots, f_*(x_m)$ . Intuitivement, cette fonction ne devrait pas être très différente de  $f_{\text{Alg}}$  mais elle ne dépend pas de  $x_i$  et  $f_*(x_i)$ . On peut donc dire que

$$\ell(f_{\text{Alg},-i}(x_i), f_*(x_i))$$

a à peu près la même distribution que  $\ell(f_{\text{Alg}}(x), f_*(x))$  pour  $x \sim \mathbb{P}$ . On a donc l'approximation suivante :

$$\text{Err}(f_{\text{Alg}}) \approx \frac{1}{m} \sum_{i=1}^m \ell(f_{\text{Alg},-i}(x_i), f_*(x_i)).$$

Avec certaines hypothèses sur l'algorithme, on peut justifier rigoureusement cette approximation [Elisseff and Pontil, 2003].

## 2 Introduction

### 2.1 Rappels

Dans les deux séances précédentes, nous avons considéré des *problèmes de reconstruction de matrices de bas rang*, c'est-à-dire des problèmes de la forme

$$\text{minimiser } \text{rang}(X) \text{ pour } X \in \mathcal{E},$$

avec  $\mathcal{E}$  un ensemble « simple ».

Si  $r$  est une borne supérieure de son rang, la solution  $X^s$  peut s'écrire sous l'une des deux formes suivantes (selon que  $X^s$  est symétrique ou non) :

$$X^s = U^t V \quad \text{ou bien} \quad X^s = U^t U.$$

Les algorithmes non-convexes consistent à reconstruire itérativement les facteurs  $(U, V)$  au moyen d'heuristiques simples. Bien qu'elles puissent a priori être bloquées par des *mauvais points critiques*, ces méthodes fonctionnent bien dans de nombreuses situations.

Dans la séance précédente, nous avons vu qu'on peut parfois montrer qu'il n'existe pas de mauvais point critique (au moins au voisinage de la solution) ; l'algorithme considéré est alors assuré de trouver la solution souhaitée.

## 2.2 Sujet de cette séance

Cette technique ne s'applique malheureusement pas à tous les algorithmes : certains fonctionnent bien alors qu'il existe de mauvais points critiques. La figure 1, par exemple, représente les bassins d'attraction de divers points critiques pour la méthode des projections alternées (utilisée ici pour la reconstruction de phase de signaux à coordonnées réelles). On constate que de mauvais points critiques existent, et même en grand nombre. Néanmoins, comme le bassin d'attraction de la solution est plus grand, en volume, que tous les bassins d'attraction des autres points réunis, la méthode des projections alternées, initialisée aléatoirement, fonctionne avec assez grande probabilité.

Estimer la taille des bassins d'attraction est en général difficile : si on note  $T$  l'opérateur qu'applique l'algorithme à chaque itération, cela nécessite de comprendre le comportement de la suite  $(T^k(x_0))_{k \in \mathbb{N}}$  en fonction de  $x_0$ . Autant il peut être possible d'analyser  $T(x_0)$ , en utilisant le fait que  $T$  est un opérateur relativement simple, aléatoire et indépendant de  $x_0$ , autant analyser  $T^2(x_0) = T(T(x_0))$  (et a fortiori  $T^k(x_0)$ ) est beaucoup plus difficile :  $T$  et  $T(x_0)$  sont aléatoires mais corrélés et la relation qui les lie est complexe.

Une méthode proposée récemment pour résoudre ce problème consiste à décorréliser partiellement  $T$  et la suite des itérées selon le principe du *tous-sauf-un*. Nous présenterons cette méthode à travers l'exemple de la synchronisation de phases, suivant l'article [Zhong and Boumal, 2018]. Cette approche est encore trop récente pour que ses extensions possibles et limitations soient bien comprises ; nous tâcherons néanmoins d'en donner une idée préliminaire en conclusion.

## 3 Synchronisation de phases

### 3.1 Définition du problème

Le problème de synchronisation de phases consiste à identifier  $n$  nombres complexes de module un,  $z_1^s, \dots, z_n^s$ , à partir de

$$c_{k,l} = z_k^s \overline{z_l^s}, \quad \forall k, l \in \{1, \dots, n\}.$$

Puisque, pour tout  $\alpha \in \mathbb{R}$ , on a  $z_k^s \overline{z_l^s} = (z_k^s e^{i\alpha}) (\overline{z_l^s e^{i\alpha}})$  pour tous  $k, l$ , l'identification n'est envisageable qu'à une phase globale près.

Si les  $(c_{k,l})_{1 \leq k, l \leq n}$  sont connus de manière exacte, la reconstruction est facile. En effet, à phase global près,

$$\begin{aligned} z_1^s &= 1; \\ z_k^s &= c_{k,1} \quad \forall k = 2, \dots, n. \end{aligned}$$

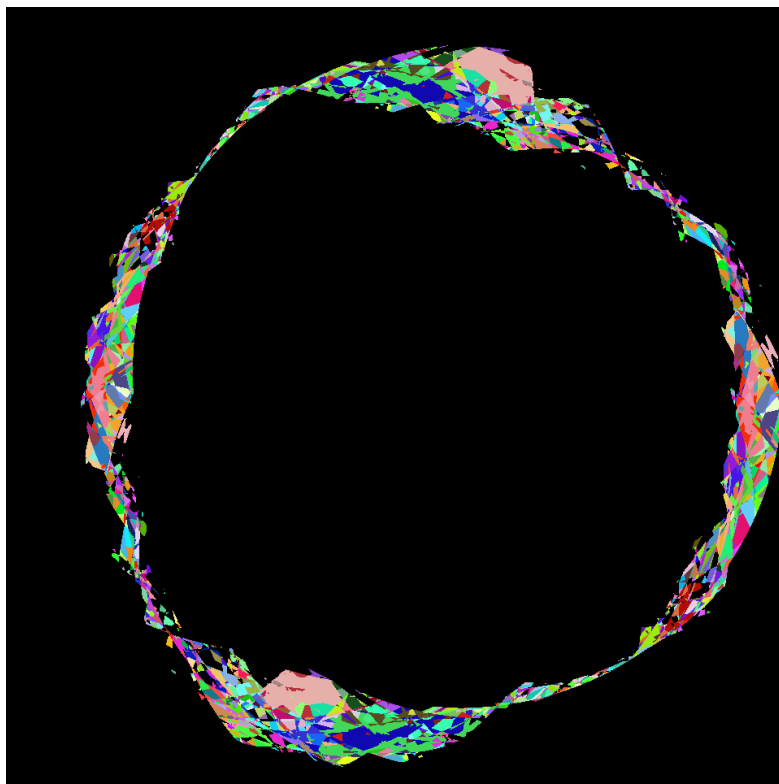


FIGURE 1 – Comportement de la méthode des projections alternées lorsqu’initialisée en chaque point d’une portion de sphère bidimensionnelle plongée dans  $\mathbb{R}^n$ . Chaque couleur représente le bassin d’attraction d’un point critique différent ; le bassin noir est celui associé à la solution. Le point au centre de la figure correspond à initialiser l’algorithme en la solution elle-même. Les points à distance 1 du centre correspondent à des points initiaux orthogonaux à la solution. Ici, la dimension est  $n = 20$  et le nombre de vecteurs de mesure est  $m = 400$ . Signaux et vecteurs de mesure sont à coordonnées réelles.

La situation qui nous intéresse ici est celle où les  $(c_{k,l})_{1 \leq k,l \leq n}$  ne sont qu'approximativement connus : pour tous  $k, l$ , le nombre qui nous est fourni n'est pas exactement  $c_{k,l}$  mais

$$C_{k,l} = z_k^s \overline{z_l^s} + w_{k,l},$$

où  $w_{k,l}$  est un « bruit ». Pour simplifier, on suppose que les  $(w_{k,l})_{1 \leq k,l \leq n}$  sont des réalisations indépendantes d'une loi normale complexe de variance  $\sigma^2$ . Plus précisément, on suppose

$$w_{k,l} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2) \text{ indépendamment, pour tous } k < l;$$

$$w_{k,l} = \overline{w_{l,k}} \text{ pour tous } k > l;$$

$$w_{k,k} = 0 \text{ pour tout } k.$$

Dans ce cas, plutôt que de reconstruire exactement les  $(z_k^s)_{1 \leq k \leq n}$  à phase globale près, on se donne pour objectif de trouver les phases  $(z_k^{obj})_{1 \leq k \leq n}$  solutions du problème de minimisation suivant :

$$\min_{|z_1|=\dots=|z_n|=1} \sum_{k,l} |C_{k,l} - z_k \overline{z_l}|^2 \quad (1)$$

(ce qui correspond à définir  $(z_k^{obj})_{1 \leq k \leq n}$  comme l'estimateur de vraisemblance maximale des  $(z_k^s)_{1 \leq k \leq n}$ ).

### 3.2 Algorithme de reconstruction et résultat principal

Remarquons que la définition (1) est équivalente à dire que  $z^{obj}$  est solution du problème de maximisation

$$z^{obj*} C z^{obj} = \max_{|z_1|=\dots=|z_n|=1} z^* C z, \quad (2)$$

ce qui est assez similaire avec la définition du vecteur propre principal de  $C$

$$z^{ppal*} C z^{ppal} = \operatorname{argmax}_{\|z\|_2=1} z^* C z.$$

Boumal [2016] a proposé de résoudre le problème (2) par l'algorithme itératif suivant, qu'il a appelé *méthode des puissances généralisée* par analogie avec la *méthode des puissances* qui permet de calculer le vecteur propre principal d'une matrice.

1. on définit  $z^{(0)} = z^{ppal}$  ;
2. pour tout  $t \geq 0$ , on pose

$$z^{(t+1)} = \mathcal{P}(C z^{(t)}),$$

où  $\mathcal{P}$  est la projection sur l'ensemble des vecteurs  $z$  tels que  $|z_1| = \dots = |z_n| = 1$ , c'est-à-dire que, pour tout  $z$ ,

$$\mathcal{P}(z)_k = \frac{z_k}{|z_k|} \quad \forall k = 1, \dots, n.$$

(On définit par convention  $\frac{0}{0} = 1$ .)

Le résultat principal de [Zhong and Boumal, 2018] est le suivant.

**Théorème 3.1.** *Il existe des constantes  $\delta > 0, \rho \in ]0; 1[$  telles que, si  $\sigma \leq \delta \sqrt{\frac{n}{\log(n)}}$ , alors, avec probabilité au moins  $1 - O\left(\frac{1}{n^2}\right)$ , la suite des itérées renvoyée par la méthode des puissances généralisées vérifie*

$$\frac{\text{dist}(z^{(t)}, z^{obj})}{\|z^{obj}\|} \leq \rho^t, \quad \forall t \geq 0,$$

où on définit  $\text{dist}(u, v) = \inf_{\alpha \in \mathbb{R}} \|u - e^{i\alpha} v\|$ .

Lorsque  $\sigma \geq \sqrt{n}$ , Gao and Zhang [2020] ont montré que  $z^{obj}$  ne permettait pas d'approcher  $z^s$  beaucoup plus précisément qu'un tirage aléatoire : pour une certaine constante  $c > 0$ ,

$$\mathbb{E} \text{dist}(z^{obj}, z^s)^2 \geq cn.$$

Par comparaison, si  $\hat{z}$  est choisi aléatoirement, avec loi uniforme, dans  $\{z, |z_1| = \dots = |z_n| = 1\}$ , on a  $\mathbb{E} \text{dist}(\hat{z}, z^s)^2 \leq 2n$ . Le problème de synchronisation de phase est donc essentiellement insoluble pour  $\sigma \geq \sqrt{n}$ ; la condition  $\sigma \leq \delta \sqrt{\frac{n}{\log(n)}}$  du théorème de [Zhong and Boumal, 2018] est ainsi assez peu restrictive.

### 3.3 Difficulté

Afin de comprendre la difficulté de ce problème et la nécessité d'introduire une technique de démonstration nouvelle, essayons de démontrer le théorème 3.1 avec la même stratégie que dans le cours précédent et voyons pourquoi elle nous mène à une impasse.

Commençons par une propriété basique mais indispensable.

**Proposition 3.2.** *Le vecteur à identifier,  $z^{obj}$ , est un point fixe de l'opérateur  $z \rightarrow \mathcal{P}(Cz)$ .*

*Démonstration.* Pour tout  $k$ ,

$$z^{obj*} C z^{obj} = \sum_{l, l' \neq k} C_{l, l'} \overline{z_l^{obj}} z_{l'}^{obj} + 2\text{Re} \left( \overline{z_k^{obj}} \sum_{l' \neq k} C_{k, l'} z_{l'}^{obj} \right) + 1.$$

Puisque  $z^{obj*} C z^{obj} = \max_{|z_1|=\dots=|z_n|=1} z^* C z$ , on doit en particulier avoir

$$\begin{aligned} & \sum_{l, l' \neq k} C_{l, l'} \overline{z_l^{obj}} z_{l'}^{obj} + 2\text{Re} \left( \overline{z_k^{obj}} \sum_{l' \neq k} C_{k, l'} z_{l'}^{obj} \right) + 1 \\ &= \max_{\theta \in \mathbb{R}} \left( \sum_{l, l' \neq k} C_{l, l'} \overline{z_l^{obj}} z_{l'}^{obj} + 2\text{Re} \left( e^{-i\theta} \sum_{l' \neq k} C_{k, l'} z_{l'}^{obj} \right) + 1 \right), \end{aligned}$$

ce qui est équivalent à ce que, pour un certain  $\lambda_k \in \mathbb{R}^+$ ,

$$\sum_{l' \neq k} C_{k,l'} z_{l'}^{obj} = \lambda_k z_k^{obj}.$$

Ainsi,

$$(Cz^{obj})_k = z_k^{obj} + \sum_{l' \neq k} C_{k,l'} z_{l'}^{obj} = (1 + \lambda_k) z_k^{obj},$$

de sorte que

$$\begin{aligned} \mathcal{P}(Cz^{obj})_k &= \mathcal{P}((1 + \lambda_k) z_k^{obj}) \\ &= z_k^{obj}. \end{aligned}$$

□

Pour suivre la stratégie de la dernière séance, il faudrait montrer que  $z \rightarrow \mathcal{P}(Cz)$  n'a pas de « mauvais point critique » dans un certain voisinage de  $z^s$  - disons pour simplifier  $B(z^s, \epsilon \|z^s\|)$ , pour un certain  $\epsilon > 0$  (très inférieur à 1 mais ne tendant pas vers 0 lorsque  $n$  tend vers l'infini). Par exemple, il faudrait montrer les deux propriétés suivantes :

1. L'opérateur  $z \rightarrow \mathcal{P}(Cz)$  est  $\rho$ -lipschitzien sur  $B(z^s, \epsilon \|z^s\|)$  pour un certain  $\rho < 1$  : pour tous  $y, y' \in B(z^s, \epsilon \|z^s\|)$ ,

$$\text{dist}(\mathcal{P}(Cy), \mathcal{P}(Cy')) \leq \rho \text{dist}(y, y'). \quad (3)$$

2. Les vecteurs  $z^{obj}, z_0$  appartiennent à  $B(z^s, \frac{\epsilon}{3} \|z^s\|)$ .

Combinées avec la proposition 3.2, ces deux propriétés permettraient de montrer que, pour tout  $t$ ,

$$\text{dist}(z^{(t)}, z^{obj}) \leq \rho^t \text{dist}(z^{(0)}, z^{obj}),$$

ce qui établirait le résultat.

Si la deuxième propriété est vraie, la première est malheureusement fausse avec grande probabilité. En effet, tout vecteur  $z \in \mathbb{C}^n$  ayant au moins une coordonnée nulle est un point de discontinuité de  $\mathcal{P}$ . Pour que la propriété (3) soit vraie, il faudrait donc que  $Cy$  n'ait aucune coordonnée nulle, quel que soit  $y \in B(z^s, \epsilon \|z^s\|)$ . Ceci peut-il se produire ?

Considérons d'abord un vecteur  $y = z^s + u$  avec  $u$  un vecteur choisi aléatoirement, avec probabilité uniforme sur  $B(0, \epsilon \|z^s\|)$ , indépendamment des bruits  $w_{k,l}$ . Rappelons que

$$C = z^s z^{s*} + W,$$

où  $W$  est la matrice dont les coefficients sont les  $w_{k,l}$ , c'est-à-dire en particulier qu'elle est hermitienne et que ses coefficients hors diagonale sont des réalisations de variables gaussiennes centrées de variance  $\sigma^2$ . Ainsi, pour tout  $k$ ,

$$\begin{aligned}(C(z^s + u))_k &= z_k^s (||z^s||^2 + \langle z^s, u \rangle) + \langle W_{:,k}, z^s + u \rangle \\ &= z_k^s (||z^s||^2 + O(\epsilon ||z^s||^2)) + O(\sigma ||z^s + u||) \\ &= nz_k^s + O(\epsilon n) + O(\sigma \sqrt{n}).\end{aligned}$$

Ici,  $W_{:,k}$  désigne la  $k$ -ième colonne de la matrice  $W$ . La deuxième égalité est due au fait que, puisque  $z^s + u$  et les coefficients de  $W$  sont indépendants,  $\langle W_{:,k}, z^s + u \rangle$  est une variable gaussienne centrée de variance  $||z^s + u||^2 \sigma^2$ ; ses valeurs sont donc d'ordre  $\sigma ||z^s + u||$ . Puisque  $|z_k^s| = 1$  et  $\sigma \ll \sqrt{n}$ , on a

$$|C(z^s + u)|_k = n(1 + O(\epsilon)),$$

ce qui est loin de 0. Ce raisonnement très informel suggère que, pour des vecteurs  $y \in B(z^s, \epsilon ||z^s||)$  choisis aléatoirement indépendamment de  $W$ ,  $Cy$  n'a pas de coordonnée nulle, ni même de coordonnée proche de zéro, avec grande probabilité.

En revanche, si on considère  $y = z^s - \eta \frac{W_{:,k}}{||W_{:,k}||}$ , pour un complexe  $\eta$  tel que  $|\eta| < \epsilon ||z^s||$  dont on précisera la valeur plus tard,

$$\begin{aligned}\left(C \left( z^s - \eta \frac{W_{:,k}}{||W_{:,k}||} \right)\right)_k &= z_k^s \left( ||z^s||^2 - \eta \left\langle z^s, \frac{W_{:,k}}{||W_{:,k}||} \right\rangle \right) + \langle W_{:,k}, z^s \rangle - \eta ||W_{:,k}|| \\ &= z_k^s n(1 + O(\epsilon)) - \eta ||W_{:,k}|| \\ &= z_k^s n(1 + O(\epsilon)) - \eta \sigma \sqrt{n}(1 + o(1)).\end{aligned}$$

En choisissant  $\eta = \frac{\sqrt{n}(1+O(\epsilon))}{\sigma(1+o(1))} z_k^s$ , on a donc

$$(Cy)_k = 0.$$

Comme  $||z^s|| = \sqrt{n}$ , on a bien  $|\eta| < \epsilon ||z^s||$  dès lors que  $\sigma > \frac{2}{\epsilon}$ . Ainsi,  $y$  est dans  $B(z^s, \epsilon ||z^s||)$  et  $Cy$  a sa  $k$ -ième coordonnée nulle.

Pour résumer, la propriété (3) est vraisemblable pour des vecteurs  $y, y'$  choisis aléatoirement dans  $B(z^s, \epsilon ||z^s||)$  indépendamment de  $W$  mais fausse lorsque  $y$  ou  $y'$  est « anormalement » aligné avec une colonne de  $W$ .

### 3.4 Démonstration

En simplifiant un peu, la démonstration du théorème 3.1 se décompose en quatre étapes.



1. On montre que l'opérateur  $z \rightarrow \mathcal{P}(Cz)$  est  $\rho$ -lipschitzien, pour un  $\rho \in ]0; 1[$ , sur

$$\mathcal{N} \stackrel{\text{déf}}{=} \{z \in B(z^s, \epsilon \|z^s\|) \mid \forall k, |\langle W_{:,k}, z \rangle| < \kappa n\},$$

avec  $\epsilon, \kappa > 0$  des constantes convenablement choisies.

2. On montre que  $z^{(0)}$  appartient, avec grande probabilité, à

$$\mathcal{N}' \stackrel{\text{déf}}{=} \left\{ z \in B\left(z^s, \frac{\epsilon}{3} \|z^s\|\right) \mid \forall k, |\langle W_{:,k}, z \rangle| < \frac{\kappa n}{2} \right\}.$$

3. On montre que  $z^{(t)}$  appartient à  $\mathcal{N}'$  pour tout  $t > 0$  avec grande probabilité.
4. On conclut : les trois étapes précédentes entraînent que  $(z^{(t)})_{t \in \mathbb{N}}$  est une suite de Cauchy. Elle converge donc, vers une limite dans  $\mathcal{N}$ . Il faut seulement vérifier que la limite est  $z^{obj}$ .

La première étape est essentiellement une formalisation du raisonnement non-rigoureux présenté dans le paragraphe précédent. La quatrième utilise une relation entre la définition de  $z^{obj}$  et un problème d'optimisation semi-définie ; nous ne la détaillerons pas. Les deuxième et troisième utilisent le principe du *tous-sauf-un* qui est l'objet de cette séance. Admettons la deuxième et concentrons-nous sur la troisième.

Lorsqu'on a montré que  $z^{(0)}, \dots, z^{(t-1)}$  appartenaient à  $\mathcal{N}'$ , on peut montrer sans difficulté majeure que

$$z^{(t)} \in B\left(z^s, \frac{\epsilon}{3} \|z^s\|\right).$$

La difficulté est de montrer que, pour tout  $k \leq n$ ,

$$|\langle W_{:,k}, z^{(t)} \rangle| < \frac{\kappa n}{2}.$$

Suivant le raisonnement du paragraphe précédent, ce serait facile si  $z^{(t)}$  et  $W_{:,k}$  étaient indépendants, pour tout  $k$ , mais  $z^{(t)}$  dépend de  $C$  et donc de  $W$ .

Pour contourner ce problème, on introduit, pour tout  $k \leq n$ , une suite auxiliaire  $(z^{(k,t)})_{t \in \mathbb{N}}$ , dont la définition est exactement la même que celle de  $z^{(t)}$ , en remplaçant la matrice

$$C = z^s z^{s*} + W$$

par la matrice

$$C^{(k)} = z^s z^{s*} + W^{(k)},$$

où  $W^{(k)}$  est la matrice  $W$  dont les coefficients des  $k$ -ième ligne et colonne ont été remplacés par zéro. Soulignons que ces suites  $(z^{(k,t)})_{t \in \mathbb{N}}$  ne pourraient pas être calculées en pratique puisqu'en pratique on ne connaît pas exactement  $W$  (sinon on connaîtrait  $z^s$ ) ; elles sont seulement des outils théoriques.

Pour tout  $k$ , la suite  $(z^{(k,t)})_{t \in \mathbb{N}}$  est indépendante de  $W_{:,k}$ . Elle vérifie donc, pour tout  $t$ , avec grande probabilité,<sup>1</sup>

$$|\langle W_{:,k}, z^{(k,t)} \rangle| = \tilde{O}(\sigma\sqrt{n}) < \frac{\kappa n}{4}.$$

En outre, on peut montrer que la suite auxiliaire  $(z^{(k,t)})_{t \in \mathbb{N}}$  et la suite principale  $(z^{(t)})_{t \in \mathbb{N}}$  sont proches ; cela permet de montrer que, pour tout  $t$ ,  $z^{(t)}$  vérifie la même propriété.

Plus précisément, on démontre par récurrence sur  $t$ <sup>2</sup> que les propriétés suivantes sont vraies avec grande probabilité :

1.  $\text{dist}(z^{(t)}, z^{(k,t)}) \leq \frac{1}{60}$  pour tout  $k$  ;
2.  $|\langle W_{:,k}, z^{(t)} \rangle| < \frac{\kappa n}{2}$  pour tout  $k$ .

Idée de démonstration pour la première propriété : pour tout  $k$ ,

$$\begin{aligned} \text{dist}(z^{(t)}, z^{(k,t)}) &= \text{dist}(\mathcal{P}(Cz^{(t-1)}), \mathcal{P}(C^{(k)}z^{(k,t-1)})) \\ &\leq \text{dist}(\mathcal{P}(Cz^{(t-1)}), \mathcal{P}(Cz^{(k,t-1)})) + \text{dist}(\mathcal{P}(Cz^{(k,t-1)}), \mathcal{P}(C^{(k)}z^{(k,t-1)})). \end{aligned} \quad (4)$$

Par hypothèse de récurrence,

$$z^{(t-1)} \in \mathcal{N}' \quad \text{et} \quad \text{dist}(z^{(k,t-1)}, z^{(t-1)}) \leq \frac{1}{60},$$

ce dont on déduit que  $z^{(t-1)}, z^{(k,t-1)}$  sont dans  $\mathcal{N}$  et donc que

$$\text{dist}(\mathcal{P}(Cz^{(t-1)}), \mathcal{P}(Cz^{(k,t-1)})) \leq \rho \text{dist}(z^{(t-1)}, z^{(k,t-1)}) \leq \frac{\rho}{60}. \quad (5)$$

D'autre part,

$$\begin{aligned} \text{dist}(\mathcal{P}(Cz^{(k,t-1)}), \mathcal{P}(C^{(k)}z^{(k,t-1)})) &\stackrel{(*)}{\leq} \frac{2}{n} \text{dist}(Cz^{(k,t-1)}, C^{(k)}z^{(k,t-1)}) \\ &\leq \frac{2}{n} \|(C - C^{(k)})z^{(k,t-1)}\| \\ &\stackrel{(\square)}{\leq} \frac{2}{n} \left( |\langle W_{:,k}, z^{(k,t-1)} \rangle| + \|W_{:,k}\| |z_k^{(k,t-1)}| \right) \\ &\stackrel{(\circ)}{=} \tilde{O}\left(\frac{\sigma}{\sqrt{n}}\right). \end{aligned} \quad (6)$$

---

1. La notation  $\tilde{O}$  est équivalente à  $O$ , à part qu'elle inclut un facteur  $\sqrt{\log(n)}$  supplémentaire.

2. En fait, pour des raisons techniques, on n'utilise la récurrence que jusqu'à  $t \approx 3n^2$ . Pour les valeurs de  $t$  supérieures à  $3n^2$ , on utilise un argument différent, plus élémentaire, que nous n'expliquerons pas.

L'inégalité (★) utilise le fait que les coordonnées de  $Cz^{(k,t-1)}$  et  $C^{(k)}z^{(k,t-1)}$  sont plus grandes, en module, que  $\frac{n}{2}$ , ce qui est une conséquence des hypothèses de récurrence. L'inégalité (□) provient du fait que  $C - C^{(k)}$  a toutes ses lignes et colonnes nulles, sauf la  $k$ -ième colonne, qui est égale à  $W_{:,k}$  et sa  $k$ -ième ligne, qui est égale à  $W_{:,k}^*$ . L'inégalité (○) utilise l'indépendance entre  $W_{:,k}$  et  $z^{(k,t-1)}$  pour majorer  $|\langle W_{:,k}, z^{(k,t-1)} \rangle|$ .

La combinaison des inégalités (4), (5) et (6) donne

$$\text{dist}(z^{(t)}, z^{(k,t)}) \leq \frac{\rho}{60} + \tilde{O}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{60}.$$

Idée de démonstration pour la deuxième propriété : pour tout  $k$ ,

$$\begin{aligned} |\langle W_{:,k}, z^{(t)} \rangle| &\leq |\langle W_{:,k}, z^{(k,t)} \rangle| + \|W_{:,k}\| \text{dist}(z^{(t)}, z^{(k,t)}) \\ &= \tilde{O}(\sigma\sqrt{n}) + \tilde{O}(\sigma\sqrt{n}) \times \frac{1}{60} \\ &< \frac{\kappa n}{2}. \end{aligned}$$

## 4 Extensions et limitations

La méthode du *tous-sauf-un* est récente en optimisation non-convexe. Sa puissance et ses limites ne sont pas encore bien établies. Nous allons néanmoins tâcher de nous en faire une idée préliminaire dans cette dernière section.

### 4.1 Principe général

Afin de comprendre comment la méthode peut être appliquée à d'autres algorithmes que celui que nous avons étudié, résumons abstraitement son principe.

Considérons un algorithme non-convexe général, dont la suite d'itérées  $(z^{(t)})_{t \in \mathbb{N}}$  est définie par

$$z^{(t+1)} = T(z^{(t)})$$

pour un certain opérateur  $T$ . Le principe de la méthode est le suivant.

1. On choisit un ensemble approprié, de la forme

$$\mathcal{N} = \{z \in \Omega \text{ tels que } F_1(z) < 1, \dots, F_K(z) < 1\},$$

pour certaines fonctions  $F_1, \dots, F_K : \mathbb{R}^n \rightarrow \mathbb{R}$  et  $\Omega$  un ouvert déterministe. (Dans le cas de la synchronisation de phases, les fonctions étaient  $F_k : z \rightarrow \frac{1}{\kappa n} |\langle W_{:,k}, z \rangle|$  pour  $k = 1, \dots, n$ ;  $\Omega$  était  $B(z^s, \epsilon \|z^s\|)$ .)

On montre que

- $T$  est contractante sur  $\mathcal{N}$  (ou une autre propriété impliquant que  $z^{(t)} \xrightarrow{t \rightarrow +\infty} z^{obj}$  si  $z^{(t)} \in \mathcal{N}$  pour tout  $t$ );
  - les fonctions  $F_1, \dots, F_K$  sont lipschitziennes au voisinage de chaque point de  $\mathcal{N}$ .
2. On montre que  $z^{(0)} \in \mathcal{N}$ .
  3. On introduit des itérées auxiliaires  $(z^{(k,t)})_{1 \leq k \leq K, t \in \mathbb{N}}$ , définies récursivement à l'aide de variantes de  $T$  :

$$z^{(k,t+1)} = T_k(z^{(k,t)}).$$

La définition de ces suites auxiliaires doit être telle que, pour tous  $k, t$ , il est possible de borner simplement

$$F_k(z^{(k,t)}).$$

4. On montre par récurrence que, pour tout  $t \geq 0$ , pour des réels  $(\epsilon_{k,t})_{1 \leq k \leq K}$  bien choisis,

$$\forall k, \quad \|z^{(t)} - z^{(k,t)}\| \leq \epsilon_{k,t};$$

$$z^{(t)} \in \mathcal{N}.$$

Le principe de démonstration de la première propriété est, dans les grandes lignes,

$$\begin{aligned} \|z^{(t)} - z^{(k,t)}\| &= \|T(z^{(t-1)}) - T_k(z^{(k,t-1)})\| \\ &\leq \|T(z^{(t-1)}) - T(z^{(k,t-1)})\| + \|T(z^{(k,t-1)}) - T_k(z^{(k,t-1)})\| \\ &\leq \rho_T \epsilon_{k,t-1} + \|(T - T_k)(z^{(k,t-1)})\|, \end{aligned}$$

où  $\rho_T$  est la constante de Lipschitz de  $T$  sur  $\mathcal{N}$ . L'opérateur  $T_k$  étant une variante de  $T$ , on peut espérer borner  $\|(T - T_k)(z^{(k,t-1)})\|$  par une quantité assez petite.

Pour la deuxième propriété, il faut montrer que  $F_k(z^{(t)}) < 1$  pour tout  $k$ . On majore

$$\begin{aligned} F_k(z^{(t)}) &\leq F_k(z^{(k,t)}) + |F_k(z^{(t)}) - F_k(z^{(k,t)})| \\ &\leq F_k(z^{(k,t)}) + \rho_{F_k} \epsilon_{k,t}, \end{aligned}$$

avec  $\rho_{F_k}$  la constante de Lipschitz de  $F_k$  au voisinage de  $z^{(t)}$  et on conclut en utilisant le fait que, grâce à la définition judicieuse de  $(z^{(k,t)})_{t \in \mathbb{N}}$ , on dispose d'une bonne borne supérieure pour  $F_k(z^{(k,t)})$ .

C'est ce schéma de démonstration qu'on retrouve, par exemple, dans [Ma, Wang, Chi, and Chen, 2018; Ding and Chen, 2020; Chen, Chi, Fan, and Ma, 2019a; Chen, Chi, Fan, Ma, and Yan, 2019b]. Il peut être raffiné. En particulier, dans [Chen, Chi, Fan, and Ma, 2019a], l'opérateur  $T$  n'est pas exactement contractant, ce qui rend difficile de contrôler  $\|T(z^{(t-1)}) - T(z^{(k,t-1)})\|$ ; les auteur-e-s y parviennent avec un deuxième argument de *tous-sauf-un*, imbriqué dans le premier.

[Zhang, 2020] peut aussi être vu comme une instanciation un peu extrême de ce schéma, pour la reconstruction de phase par projections alternées dans le cas où le nombre de vecteurs de mesures est beaucoup plus grand que la dimension. Dans cet article,  $T_k$  n’est pas une version légèrement modifiée de  $T$  ; c’est un opérateur aléatoire de même distribution que  $T$  mais indépendant.

## 4.2 Difficultés

Un premier inconvénient de cette méthode est qu’elle requiert une compréhension assez fine des propriétés de  $T$  et de  $dT$ , plus particulièrement de leur lipschitzianité locale. Elle aboutit donc à des démonstrations assez techniques et spécifiques à un problème et un algorithme.

Il me semble qu’un deuxième inconvénient est qu’elle repose fortement sur le fait que  $T$  est continu et même contractant sur une grande partie de son espace de définition. Cela permet de garantir que les suites  $(z^{(t)})_{t \in \mathbb{N}}$  et  $(z^{(k,t)})_{t \in \mathbb{N}}$  restent proches l’une de l’autre. Or certains algorithmes ne vérifient pas cette propriété, par exemple la méthode des projections alternées lorsqu’elle est appliquée à des signaux et vecteurs de mesure à coordonnées réelles, ces derniers suivant une loi normale<sup>3</sup>. Un calcul sommaire suggère en effet que, dans ce cas, pour tous points  $z, z'$  sur la sphère unité,

$$\mathbb{E}(\|T(z) - T(z')\|^2) \text{ est de l'ordre de } \frac{n}{m} \|z - z'\|,$$

ce qui suggère que  $T$  est loin d’être localement contractant. La figure 2 confirme ce fait : pour l’algorithme *Wirtinger Flow*, décrit à la séance précédente, auquel la méthode du *tous-sauf-un* s’applique [Chen, Chi, Fan, and Ma, 2019a],  $\|T(z) - T(z')\|$  est à peu près proportionnelle à  $\|z - z'\|$ , avec un coefficient de proportionnalité inférieur à 1. Pour la méthode des projections alternées,  $\|T(z) - T(z')\|$  croît beaucoup plus rapidement.

## Références

- N. Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4) :2355–2377, 2016.
- Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization : Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2) :5–37, 2019a.

---

3. Sur ce point précis, le comportement de la méthode dans le cas réel semble très différent du cas complexe.

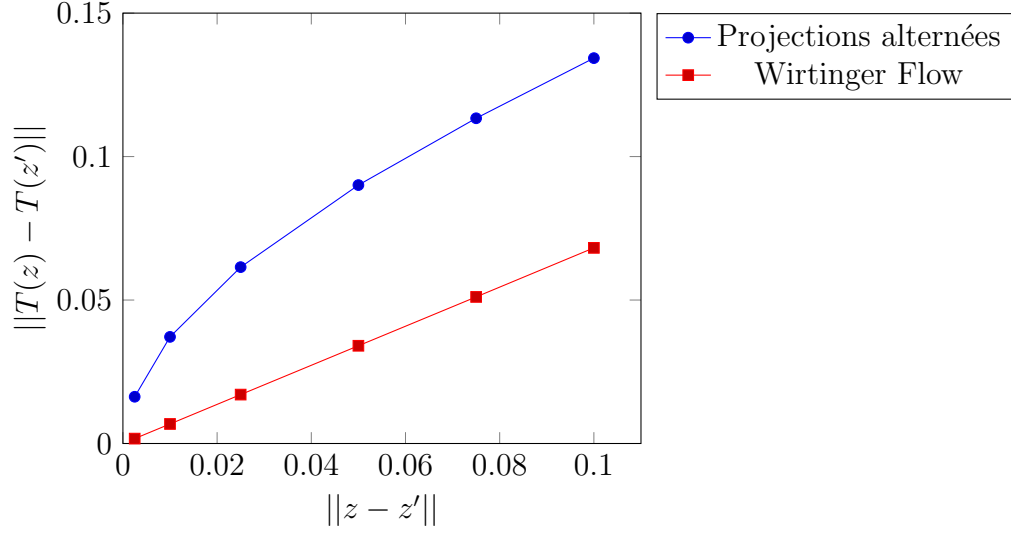


FIGURE 2 – Évolution de  $\|T(z) - T(z')\|$  en fonction de  $\|z - z'\|$ , moyennée sur 1000 choix de paires  $(z, z')$ , pour deux algorithmes reconstruisant des signaux de dimension  $n = 400$  à partir de  $m = 4000$  mesures sans phase.

- Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion : Understanding statistical guarantees for convex relaxation via nonconvex optimization. preprint, 2019b. <https://arxiv.org/abs/1902.07698>.
- L. Ding and Y. Chen. Leave-one-out approach for matrix completion : Primal and dual analysis. IEEE Transactions on Information Theory, 2020.
- N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. Proceedings of the National Academy of Sciences, 110(36) :14557–14562, 2013.
- A. Elisseeff and M. Pontil. Leave-one-out error and stability of learning algorithms with applications. In NATO science series, III : Computer and systems sciences, volume 190, pages 111–130. IOS Press, 2003.
- C. Gao and A. Y. Zhang. Exact minimax estimation for phase synchronization. preprint, 2020. <https://arxiv.org/abs/2010.04345>.
- C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation : Gradient descent converges linearly for phase retrieval and matrix completion. In International Conference on Machine Learning, pages 3345–3354, 2018.

- T. Zhang. Phase retrieval by alternating minimization with random initialization. IEEE Transactions on Information Theory, 2020.
- Y. Zhong and N. Boumal. Near-optimal bounds for phase synchronization. SIAM Journal on Optimization, 28(2) :989–1016, 2018.