Méthodes non-convexes sans mauvais point critique

Jeudi 5 novembre 2020

1 Rappels de la première séance

On appelle *problème de reconstruction de matrice de bas rang* un problème de la forme

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

où $\mathcal{E}$ est un sous-ensemble « simple » de $\mathrm{Mat}(n_1, n_2)$, avec n_1, n_2 des entiers fixés..

Nous avons vu plusieurs exemples de tels problèmes. La *reconstruction de phase*, notamment, consiste à identifier (à phase globale près) un vecteur $x^s \in \mathbb{C}^n$ à partir de

$$b_1 = |\langle x^s, v_1\rangle|, \dots, b_m = |\langle x^s, v_m\rangle|,$$

où $v_1, \dots, v_m$ sont des vecteurs de mesure fixés. Bien que ce problème consiste a priori à reconstruire un vecteur et non une matrice, on peut le voir comme un problème de reconstruction de matrice de bas rang au moyen du changement de variable $X^s = x^s(x^s)^*$.

Diverses heuristiques naturelles, dites *méthodes non-convexes*, ont été développées pour résoudre ces problèmes mais ne fonctionnent pas à coup sûr : elles peuvent restées bloquées dans un optimum local. De plus, ces méthodes sont difficiles à analyser théoriquement et on ne peut pas déterminer à l'avance dans quelles situations elles fonctionneront correctement.

À la séance précédente, nous avions expliqué que les *méthodes convexes* permettaient en partie de contourner ces difficultés :

— elles peuvent fonctionner dans certaines situations où les méthodes non-convexes sont victimes de la présence d'optima locaux ;

— leur analyse théorique est un peu plus facile.

Dans le cas de la reconstruction de phase, la méthode convexe la plus connue, est *PhaseLift*, qui admet les garanties de correction suivantes :

Théorème 1.1 ([Candès and Li, 2014])**.** *Supposons que les vecteurs de mesure $v_1, \dots, v_m$ sont des réalisations indépendantes d'une loi normale $\mathcal{N}(0, I_n)$.*

Il existe des constantes $C, c > 0$ telles que, pour tous $n, m \in \mathbb{N}^*$ vérifiant $m \geq Cn$, avec probabilité au moins $1 - e^{-cm}$, PhaseLift *résout correctement le problème de reconstruction de phase.*

Malheureusement, les méthodes convexes sont difficilement applicables en pratique car très coûteuses en temps de calcul, ce qui a provoqué ces dernières années un regain d'intérêt pour les méthodes non-convexes.

2 Introduction

2.1 Définition informelle des méthodes non-convexes

Une matrice $X \in \mathrm{Mat}(n_1, n_2)$ de rang r peut s'écrire sous la forme

$$X = UV^T, \qquad \text{avec } U \in \mathrm{Mat}(n_1, r), V \in \mathrm{Mat}(n_2, r).$$

Lorsque X est symétrique, on peut même écrire $X = UU^T$. Reconstruire U et V suffit à reconstruire X.

Remarque 2.1. *Dans le cas particulier de la reconstruction de phase, cette décomposition est évidente puisque la matrice inconnue est, par définition, $X^s = x^s(x^s)^*$, avec x^s le signal à reconstruire.*

Les algorithmes non-convexes fonctionnent en général selon le schéma suivant :

1. Choix d'un point initial (U_0, V_0) (qui peut être une estimation de la solution ou un point arbitraire).
2. Application récursive d'une heuristique simple visant à rapprocher (U_0, V_0) de la solution cherchée. Cela donne une suite d'itérées $(U_t, V_t)_{t\in\mathbb{N}}$ qui, idéalement, converge vers une solution.
3. Renvoi de $U_T V_T^T$ pour un $T \in \mathbb{N}$ assez grand.

Ce sont les méthodes les plus utilisées en pratique. Si, comme nous l'avons vu, elles peuvent échouer, on constate empiriquement qu'elles fonctionnent bien dans un certain nombre de cas.

2.2 Un exemple : reconstruction de phase par projections alternées

Pour donner un exemple d'un tel cas, considérons à nouveau un problème de reconstruction de phase pour des vecteurs de mesure générés aléatoirement selon une loi normale :

$$v_1, \ldots, v_m \overset{iid}{\sim} \mathcal{N}(0, I_n).$$

Étudions numériquement la performance de l'algorithme non-convexe le plus ancien et le plus classique, la méthode des projections alternées (aussi appelé *error reduction* ou *Gerchberg-Saxton*, du nom des chercheurs l'ayant introduit [Gerchberg and Saxton, 1972]).

Pour définir la méthode des projections alternées, introduisons, pour tout $x \in \mathbb{C}^n$, la notation $\mathcal{A}(x) = (\langle x, v_1 \rangle, \dots, \langle x, v_m \rangle)$. Lorsque $m \geq n$, $\mathcal{A}$ est un opérateur linéaire injectif avec probabilité 1. Retrouver le vecteur x^s (à phase globale près) est donc équivalent à reconstruire $y^s \stackrel{def}{=} \mathcal{A}(x^s)$ (à phase globale près). Le problème de reconstruction de phase peut donc être reformulé de la manière suivante :

$$\begin{aligned} &\text{trouver } y \in \mathbb{C}^m, \\ &\text{tel que } y \in \mathrm{Im}(\mathcal{A}), \\ &\qquad |y_k| = b_k, \quad \forall k \leq m. \end{aligned}$$

En définissant $E = \{y \in \mathbb{C}^m, |y_k| = b_k\}$, ce problème peut s'écrire plus concisément :

$$\begin{aligned} &\text{trouver } y \in \mathbb{C}^m, \\ &\text{tel que } y \in \mathrm{Im}(\mathcal{A}) \cap E. \end{aligned}$$

Définition 2.2. *Étant fixé $T \in \mathbb{N}$, la méthode des projections alternées consiste à :*

1. *choisir un point de départ y_0 arbitraire ;*
   (Ici, nous choisirons $y_0 \sim \mathcal{N}(0, I_m)$.)
2. *pour tout $t \in \{1, \dots, T\}$, définir*

$$y_t = P_{\mathrm{Im}(\mathcal{A})} P_E(y_{t-1}),$$

   où, pour tout ensemble fermé S, P_S désigne la projection sur S [1] *;*
3. *renvoyer y_T.*

La figure 1 [2] montre la courbe de performance de la méthode des projections alternées (mesurée en termes de probabilité de reconstruction exacte), en fonction de m/n, pour $n = 40$. Au vu de cette figure, il est tentant de conjecturer que, lorsque les vecteurs de mesure ont une loi normale, la méthode des projections alternées fonctionne avec grande probabilité dès lors que $m \geq Cn$ pour une certaine constante $C > 0$, comme *PhaseLift*. Peut-on le démontrer ?

1. c'est-à-dire une application $P_S : \mathbb{C}^m \to S$ telle que, pour tout y, $||P_S(y) - y|| = \min_{z \in S} ||z - y||$

2. La courbe rouge de cette figure a été générée à l'aide de la librairie *PhasePack* [Chandra, Zhong, Hontz, McCulloch, Studer, and Goldstein, 2017]. Quant à la courbe bleue et à toutes les autres figures contenues dans ces notes de cours, elles ont été générées avec le code disponible à l'adresse `https://www.ceremade.dauphine.fr/~waldspurger/code/non_convex_lecture_notes_figures.zip`.

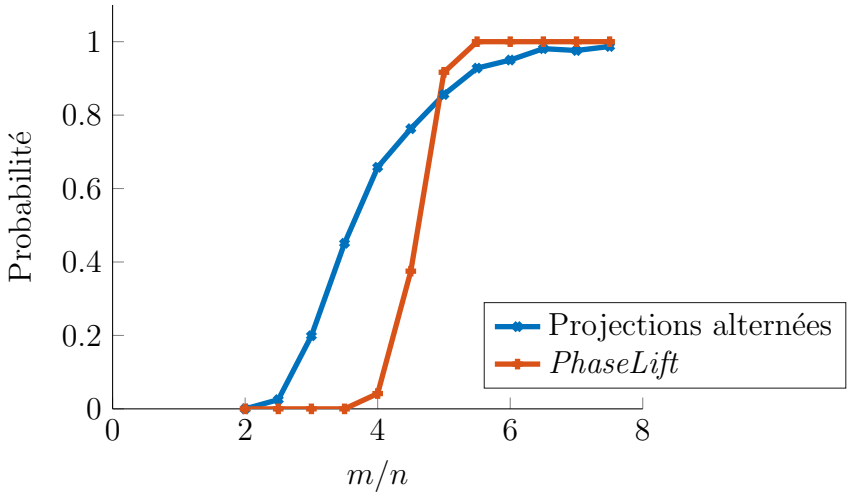

FIGURE 1 – Probabilité de succès pour deux algorithmes de reconstruction de phase, en fonction de m/n, pour $n = 40$.

2.3 Plan de la séance

La principale difficulté dans l'analyse des algorithmes non-convexes est la présence possible de points de stagnation, qu'on appellera dans la suite des *mauvais points critiques* : à l'étape t, si l'heuristique échoue à rapprocher x_t de la solution car x_t vérifie une sorte de propriété d'optimalité locale, il est possible que la suite des itérées reste au voisinage de x_t bien que x_t ne soit pas la solution.

Pour démontrer qu'un algorithme non-convexe fonctionne, une stratégie de preuve naturelle consiste donc à démontrer (si c'est vrai) qu'il n'existe pas de points de stagnation. Cette stratégie est inapplicable pour beaucoup d'algorithmes naturels. La méthode des projections alternées, par exemple, semble avoir des points de stagnation même lorsque m/n est très grand. Néanmoins, il est possible de concevoir des algorithmes non-convexes un peu plus sophistiqués pour lesquels cette stratégie convient. Cet axe de recherche est le sujet de cette séance.

La stratégie de démonstration peut être mis en œuvre de deux manières différentes.

— On peut

1. démontrer d'abord que l'heuristique utilisée par l'algorithme n'a pas de mauvais point critique *dans un certain voisinage de la solution* ;
2. puis montrer que toutes les itérées appartiennent à ce voisinage de la solution (ce qui nécessite en particulier que l'algorithme utilise une stratégie d'initialisation subtile).

Nous décrirons cette approche dans la section 3.

— On peut démontrer que l'heuristique n'a pas de mauvais point critique du tout. C'est l'objet de la section 4.

Nous décrirons ces méthodes dans le cadre particulier de la reconstruction de phase, pour des vecteurs de mesure distribués selon des lois normales. Elles s'appliquent à de nombreux autres problèmes ; des références sont données à la fin des sections 3 et 4.

3 Pas de mauvais point critique près de la solution

Nous allons présenter la première des deux méthodes à travers l'algorithme *Wirtinger Flow*, introduit dans [Candès, Li, and Soltanolkotabi, 2015], dont le fonctionnement est le suivant :

1. Initialisation : choix de x_0 au moyen d'une méthode dite *spectrale* que nous décrirons par la suite ;
2. Raffinement : soit
$$f : x \in \mathbb{C}^n \to \frac{1}{2m} \sum_{k=1}^{m} (|\langle x, v_k \rangle|^2 - b_k^2)^2.$$
C'est une fonction $\mathcal{C}^\infty$, non-convexe, dont les minima sont exactement les solutions du problème de reconstruction de phase [3].
Pour tout $t \in \mathbb{N}$, on définit x_{t+1} en appliquant une étape de descente de gradient à f, à partir de x_t :
$$x_{t+1} = x_t - \mu \nabla f(x_t),$$
pour une certaine constante $\mu > 0$.
3. Renvoi de x_T pour T assez grand.

Cet algorithme admet les garanties décrites dans le théorème suivant [4].

Théorème 3.1 ([Candès, Li, and Soltanolkotabi, 2015, Thm 3.3]). *Soit x^s un vecteur arbitraire. Il existe $C, c > 0$ des constantes telles que, si*

$$Cn \log(n) \leq m$$

et si $\mu \leq \frac{c}{n}$, alors, avec probabilité $1 - O\left(\frac{1}{n^2}\right)$,

$$\text{dist}(x^s, x_t) \leq \frac{1}{8} \left(1 - \frac{\mu}{4}\right)^{t/2} ||x^s||$$

pour tout $t \in \mathbb{N}$. En particulier, $x_t \overset{t \to +\infty}{\to} x^s$.

3. pour tout x, $f(x) \geq 0$ avec égalité si et seulement si $|\langle x, v_k \rangle| = b_k$ pour tout k

4. Dans ce théorème, la notation dist est définie par $\text{dist}(x, y) = \min_{\phi \in \mathbb{R}} ||x - e^{i\phi} y||$. Dans la suite des explications, on fera comme si $\text{dist}(x, y) = ||x - y||$ pour simplifier mais cela n'est pas parfaitement rigoureux.

Le but de cette section est de présenter le principe (un peu simplifié) de la preuve de ce théorème qui, comme dit précédemment, consiste en deux étapes :

1. On montre que l'heuristique utilisée à l'étape de raffinement n'a "pas de mauvais point critique" dans la boule $B(x^s, ||x^s||/8)$. Plus précisément, on établit que, pour tout $x \in B(x^s, ||x^s||/8)$,
$$\text{dist}(x^s, x - \mu\nabla f(x)) \leq \rho \text{dist}(x^s, x) \tag{1}$$
pour un $\rho = \sqrt{1 - \frac{\mu}{4}} \in]0; 1[$.
2. On analyse la méthode d'initialisation spectrale pour montrer que
$$x_0 \in B(x^s, ||x^s||/8).$$

3.1 Première étape : absence de mauvais point critique

Supposons pour simplifier $||x^s|| = 1$.

Nous allons voir comment établir la propriété (1), mais pour un $\rho \in]0; 1[$ différent de $\sqrt{1 - \frac{\mu}{4}}$, ce qui est moins optimal du point de vue de la vitesse de convergence de $(x_t)_{t\in\mathbb{N}}$ vers x^s mais permet une démonstration un peu plus simple.

On peut se contenter d'établir les propriétés suivantes :

$$\forall x \in B(x^s, 1/8), \quad \text{Re}(\langle x - x^s, \nabla f(x)\rangle) \geq \alpha \text{dist}(x, x^s)^2, \tag{2}$$
$$\forall x \in B(x^s, 1/8), \quad ||\nabla f(x)|| \leq \beta \text{dist}(x, x^s), \tag{3}$$

pour $\alpha, \beta > 0$ des valeurs bien choisies. En effet, si ces propriétés sont vraies, on a, pour tout $x \in B(x^s, 1/8)$,

$$\begin{aligned} ||x^s - (x - \mu\nabla f(x))||^2 &= ||x^s - x||^2 - 2\mu \text{Re}(\langle x - x^s, \nabla f(x)\rangle) + \mu^2 ||\nabla f(x)||^2 \\ &\leq ||x^s - x||^2 - 2\mu\alpha ||x - x^s||^2 + \mu^2\beta^2 ||x - x^s||^2 \\ &\leq (1 - \mu\alpha)||x^s - x||^2 \quad \text{si } \mu < \frac{\alpha}{\beta^2}. \end{aligned}$$

Le principe général de la démonstration des propriétés (2) et (3) est d'utiliser l'expression explicite du gradient pour écrire $\text{Re}(\langle x - x^s, \nabla f(x)\rangle)$ et $||\nabla f(x)||^2$ comme une somme de réalisations de variables aléatoires indépendantes. Cette somme peut être analysée à l'aide d'outils statistiques classiques : les inégalités de concentration. Expliquons brièvement ce que sont ces inégalités, pour les lecteur/trice/s qui n'en seraient pas familier/ère/s : dans leur version la plus basique, le but des inégalités de concentration est de décrire le comportement d'une somme de variables aléatoires indépendantes

$$Y_1 + Y_2 + \cdots + Y_K.$$

Sous des hypothèses raisonnables, la somme est « proche » de son espérance avec grande probabilité lorsque K est assez grand ; lorsque les Y_k sont de même distribution, c'est la loi des grands nombres. Les inégalités de concentration permettent de contrôler précisément cette probabilité, en fournissant des majorations pour

$$\text{Proba}\left(Y_1 + \cdots + Y_K \geq \mathbb{E}(Y_1 + \cdots + Y_K) + \epsilon\right)$$

pour tout $\epsilon > 0$. La forme exacte de la majoration dépend des hypothèses qu'on peut faire sur les variables aléatoires Y_k.

Concentrons-nous sur la propriété (2). Considérons x de la forme $x = x^s + h$ avec $||h|| < 1/8$. On a

$$\begin{aligned}\nabla f(x) &= \frac{1}{m}\sum_{r=1}^{m}(|\langle v_r, x\rangle|^2 - b_r^2)v_r v_r^* x \\ &= \frac{1}{m}\sum_{r=1}^{m}\left(2\text{Re}(\overline{\langle v_r, x^s\rangle}\langle v_r, h\rangle) + |\langle v_r, h\rangle|^2\right)\langle v_r, x^s + h\rangle v_r\end{aligned}$$

et on calcule que

$$\begin{aligned}&\text{Re}(\langle x - x^s, \nabla f(x)\rangle) \\ &= \frac{1}{m}\sum_{r=1}^{m}\left(2\text{Re}^2(\overline{\langle v_r, x^s\rangle}\langle v_r, h\rangle) + 3\text{Re}(\overline{\langle v_r, x^s\rangle}\langle v_r, h\rangle)|\langle v_r, h\rangle|^2 + |\langle v_r, h\rangle|^4\right) \\ &\overset{\text{déf}}{=} \frac{1}{m}\sum_{r=1}^{m} Y_r(h).\end{aligned}$$

Pour tout h fixé, les variables aléatoires $Y_1(h), \ldots, Y_r(h)$ sont indépendantes de même loi et on peut vérifier que, pour tout r, en supposant pour simplifier $\langle x^s, h\rangle \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}(Y_r(h)) &= 3\langle x^s, h\rangle^2 + ||h||^2 + 6\langle x^s, h\rangle||h||^2 + 2||h||^4 \\ &\geq \frac{||h||^2}{2} \text{ si } ||h|| \leq \frac{1}{8}.\end{aligned}$$

Les inégalités de concentration évoquées plus haut permettent de montrer que

$$\frac{1}{m}\sum_{r=1}^{m} Y_r(h) \geq \mathbb{E}\left(\frac{1}{m}\sum_{r=1}^{m} Y_r(h)\right) - \frac{||h||^2}{4}$$

avec probabilité au moins $1 - e^{-\gamma m}$ pour une certaine constante $\gamma > 0$. On en déduit

$$\text{Re}(\langle x - x^s, \nabla f(x)\rangle) \geq \frac{||h||^2}{4} = \frac{||x - x^s||^2}{4}.$$

Cette démonstration n'est valable que pour un $x \in B(x^s, 1/8)$ fixé et non pour tous les éléments $x \in B(x^s, 1/8)$ mais on peut l'étendre à tous les éléments de la boule avec un argument de probabilité très classique (dit ϵ-*net*).

Si certaines personnes veulent plus de détails sur l'application des inégalités de concentration : la présence dans $Y_r(h)$ de puissances troisième et quatrième de $\langle v_r, h\rangle$ impose un certain soin. On peut décomposer $Y_r(h) = Y_r^{(1)}(h) + Y_r^{(2)}(h)$, avec

$$Y_r^{(1)}(h) = -\frac{1}{4}\mathrm{Re}^2(\overline{\langle v_r, x^s\rangle}\,\langle v_r, h\rangle);$$
$$Y_r^{(2)}(h) = \left(\frac{3}{2}\mathrm{Re}(\overline{\langle v_r, x^s\rangle}\,\langle v_r, h\rangle) + |\,\langle v_r, h\rangle\,|^2\right)^2.$$

Sur un événement de probabilité au moins $1 - O\left(\frac{1}{n^2}\right)$,

$$\frac{1}{m}\sum_{r=1}^{m} Y_r^{(1)}(h) \geq \mathbb{E}(Y_r^{(1)}(h)) - \eta||h||^2$$

pour tous les $h \in \mathbb{C}^n$, où $\eta > 0$ est une constante qu'on peut choisir arbitrairement petite.

Pour la somme des $Y_r^{(2)}(h)$, on utilise le lemme de concentration suivant.

Lemme 3.2 (Conséquence de [Candès, Li, and Soltanolkotabi, 2015, Lemme 7.13])**.** *Soient $Z_1, \ldots, Z_m$ des variables aléatoires indépendantes et identiquement distribuées, telles que $Z_r \geq 0$ presque sûrement, pour tout r. Alors, pour tout $z \geq 0$,*

$$P\left(\frac{1}{m}\sum_{r=1}^{m} Z_r \leq \mathbb{E}(Z_r) - z\right) \leq \exp\left(-\frac{mz^2}{2\max(\mathbb{E}(Z_1)^2, \mathrm{Var}(Z_1))}\right).$$

On obtient que, pour h fixé,

$$\frac{1}{m}\sum_{r=1}^{m} Y_r^{(2)}(h) \geq \mathbb{E}(Y_r^{(2)}(h)) - \eta||h||^2$$

avec probabilité au moins $1 - e^{-\gamma m}$ pour un certain $\gamma > 0$.

3.2 Deuxième étape : initialisation spectrale

Commençons par décrire la méthode d'initialisation spectrale. À ma connaissance, cette méthode a été proposée pour la première fois en reconstruction de phase dans [Netrapalli, Jain, and Sanghavi, 2013], avant d'être utilisée dans [Candès, Li, and Soltanolkotabi, 2015]. Une idée

similaire était déjà présente dans [Keshavan, Montanari, and Oh, 2010] mais appliquée à des problèmes de complétion de matrices.

Définissons la matrice

$$M = \frac{1}{m}\sum_{r=1}^{m} b_r^2 v_r v_r^* = \frac{1}{m}\sum_{r=1}^{m} |\langle x^s, v_r\rangle|^2 v_r v_r^* \in \mathbb{C}^{n\times n}.$$

Informellement, dans cette définition, le terme $v_r v_r^*$ apparaît avec une pondération égale à $|\langle x^s, v_r\rangle|^2$; il a donc d'autant plus d'influence sur la somme qu'il est aligné avec $x^s(x^s)^*$. Ainsi, M est « biaisée » dans la direction de $x^s(x^s)^*$ et on peut utiliser comme point initial

$$x_0 = \text{vecteur propre principal}(M). \tag{4}$$

Le lemme qui suit garantit la précision de la méthode d'initialisation spectrale.

Lemme 3.3. *Il existe une constante $C > 0$ telle que, lorsque $m \geq Cn\log(n)$,*

$$\text{dist}(x_0, x^s) \leq \frac{||x^s||}{8}$$

avec probabilité au moins $1 - O\left(\frac{1}{n^2}\right)$.

La démonstration de ce lemme passe par la propriété, valable avec probabilité $1 - O\left(\frac{1}{n^2}\right)$,

$$|||M - \mathbb{E}(M)||| \leq \delta,$$

où δ est une constante qui peut être choisie arbitrairement petite pourvu que la constante C du lemme soit assez grande. Cette propriété se démontre avec des inégalités de concentration, comme celles que nous avons évoquées dans le paragraphe précédent. Elle permet de conclure grâce à l'égalité

$$\mathbb{E}(M) = I_n + x^s(x^s)^*.$$

3.3 Autres travaux

Le principe « initialisation spectrale + raffinement », ainsi que les techniques de démonstration que nous venons de présenter, ne sont pas spécifiques à *Wirtinger Flow* ou à la reconstruction de phase. De nombreux autres algorithmes reposant sur le même principe ont été proposés et analysés.

Dans le cas de la reconstruction de phase, on peut par exemple obtenir des garanties de correction similaires au théorème 3.1 (en fait, un peu meilleures) en remplaçant l'initialisation spectrale du paragraphe précédent par des variantes un peu plus sophistiquées [Chen and Candès,

2017; Mondelli and Montanari, 2017] et en utilisant une autre heuristique pour le raffinement, par exemple une descente de gradient « tronquée » [Chen and Candès, 2017], une descente de gradient sur une autre fonction de coût (éventuellement non $\mathcal{C}^\infty$, ce qui pose quelques difficultés techniques) [Zhang and Liang, 2016; Wang, Giannakis, and Eldar, 2017] ou la méthode des projections alternées [Waldspurger, 2018].

En-dehors de la reconstruction de phase, citons par exemple [Jain, Netrapalli, and Sanghavi, 2013] pour la complétion de matrices et le *matrix sensing RIP* (reconstruction de matrices à partir de mesures linéaires lorsque l'opérateur de mesure vérifie une propriété d'isométrie restreinte), [Zhao, Wang, and Liu, 2015], pour le *matrix sensing RIP* également mais pour une plus large gamme d'heuristiques, [Zheng and Lafferty, 2016], pour la complétion de matrices à nouveau, et [Chen and Wainwright, 2015] pour des résultats plus généraux, applicables à plusieurs problèmes différents.

4 Pas de mauvais point critique du tout

La deuxième méthode que nous allons décrire permettant de démontrer la correction d'un algorithme non-convexe consiste à démontrer que cet algorithme n'a pas de mauvais point critique du tout. D'un point de vue technique, elle est en général plus compliquée à mettre en œuvre que la première méthode car elle nécessite une analyse encore plus fine des fonctions en jeu. Elle a en revanche l'avantage de s'appliquer à des algorithmes conceptuellement un peu plus simples et plus proches de la pratique car ne comportant pas de procédure d'initialisation sophistiquée.

L'algorithme à travers lequel nous illustrerons cette méthode est celui de [Sun, Qu, and Wright, 2017a]. Il est identique à *Wirtinger Flow*, à part pour l'initialisation :

1. Initialisation : choix de x_0 aléatoirement, avec probabilité uniforme sur $B(0,1)$ (par exemple).
2. Raffinement : descente de gradient [5] sur la fonction
$$f : x \in \mathbb{C}^n \to \frac{1}{2m}\sum_{k=1}^{m}(|\langle x, v_k\rangle|^2 - b_k^2)^2.$$
3. Renvoi de x_T pour T assez grand.

Les garanties de correction que nous allons établir à son sujet sont les suivantes.

5. Sun, Qu, and Wright [2017a] proposent de raffiner l'estimation x_0 en appliquant à f non pas une descente de gradient mais une autre méthode d'optimisation locale, *Trust-region*. Le théorème de convergence donné ci-après étant valable pour les deux méthodes, nous utilisons plutôt la descente de gradient.

Théorème 4.1 ([Sun, Qu, and Wright, 2017a]). *Soit x^s un vecteur arbitraire. Il existe $C > 0$ une constante telle que, si*

$$Cn\log^3(n) \leq m,$$

alors, avec probabilité $1 - O\left(\frac{1}{n}\right)$, la suite d'itérées $(x_t)_{t\in\mathbb{N}}$ calculée par l'algorithme vérifie

$$\operatorname{dist}(x^s, x_t) \overset{t\to+\infty}{\to} 0,$$

pourvu que le pas de la descente de gradient soit assez petit.

4.1 Propriétés des point critiques ?

Pour analyser l'algorithme, il faut commencer par comprendre les propriétés des points que nous appelons "critiques", c'est-à-dire ceux qui peuvent provoquer une stagnation de la descente de gradient. Nous pourrons ensuite montrer qu'il n'existe pas de point vérifiant ces propriétés, c'est-à-dire qu'il n'y a pas de mauvais point critique.

On voit immédiatement, à partir de la définition de la descente de gradient, qu'un point critique x_* vérifie nécessairement une condition d'*optimalité d'ordre* 1 :

$$\nabla f(x_*) = 0. \tag{5}$$

En outre, si on exclut un ensemble de points initiaux x_0 de mesure de Lebesgue nulle, on peut montrer (mais c'est beaucoup moins immédiat) que les points critiques vérifient également une condition d'*optimalité d'ordre* 2 :

$$\nabla^2 f(x_*) \succeq 0. \tag{6}$$

Ces propriétés, qui ne sont pas spécifiques à la fonction f que nous considérons, sont énoncées précisément dans le théorème suivant.

Théorème 4.2. *Soit $\mathcal{L}$ une fonction analytique telle que*

$$\mathcal{L}(x) \overset{||x||\to+\infty}{\to} +\infty.$$

On lui applique une descente de gradient avec pas constant $\mu > 0$, ce qui donne une suite d'itérées $(x_t)_{t\in\mathbb{N}}$. Pourvu que μ soit suffisamment petit, $(x_t)_{t\in\mathbb{N}}$ converge quel que soit $x_0 \in B(0, 1)$.

De plus, la limite x_*

- *vérifie $\nabla\mathcal{L}(x_*) = 0$ quel que soit x_0 ;*
- *vérifie $\nabla^2\mathcal{L}(x_*) \succeq 0$ pour presque tout x_0.*

La première partie de ce théorème se déduit de [Absil, Mahony, and Andrews, 2005, Thm 4.1] et la seconde de [Panageas and Piliouras, 2016, Thm 3], qui est une généralisation de [Lee, Simchowitz, Jordan, and Recht, 2016, Corollary 9].

Remarque 4.3. *Le théorème précédent est vrai pour beaucoup d'algorithmes d'optimisation locale outre la descente de gradient à pas constant.*

Pour démontrer le théorème 4.1, il suffit donc de montrer qu'il n'existe aucun point autre que la solution x^s qui vérifie les conditions d'optimalité d'ordre 1 et 2 (équations (5) et (6)). Le but de la sous-section suivante est d'expliquer comment démontrer cette non-existence.

4.2 Idée de démonstration

Pour se former une intuition de la démonstration, déterminons d'abord les points critiques de l'espérance de f ; c'est plus facile que ceux de f. Pour tout $x \in \mathbb{C}^n$ fixé, on vérifie que

$$\mathbb{E}(f(x)) = ||x||^4 - ||x||^2||x^s||^2 - |\langle x, x^s\rangle|^2 + ||x^s||^4.$$

On a donc

$$\nabla(\mathbb{E}f)(x) = 2((2||x||^2 - ||x^s||^2)x - \langle x^s, x\rangle x^s),$$

ce dont on déduit que les points satisfaisant la condition d'optimalité d'ordre 1 sont tous ceux contenus dans les trois ensembles suivants :

- $E_1 = \{e^{i\theta}x^s, \theta \in \mathbb{R}\}$, c'est-à-dire l'ensemble des solutions ;
- $E_2 = \{0\}$;
- $E_3 = \left\{x \in \mathbb{C}^n, \langle x^s, x\rangle = 0, ||x|| = \frac{||x^s||}{\sqrt{2}}\right\}$.

Parmi ces points, lesquels vérifient la condition d'optimalité d'ordre 2 ? Les points de E_1 la vérifient car ils sont les minima de $\mathbb{E}(f)$. Pour les autres, nous avons besoin de l'expression explicite de la dérivée seconde de $\mathbb{E}f$, qui est, en tout point x,

$$\nabla^2(\mathbb{E}f)(x)\cdot(h,h) = 2\left((2||x||^2 - ||x^s||^2)||h||^2 + 4\mathrm{Re}^2(\langle x, h\rangle) - |\langle x^s, h\rangle|^2\right).$$

À partir de cette expression, il est facile de voir que le point 0 ne vérifie pas la condition d'optimalité d'ordre 2 (on a même $\nabla^2(\mathbb{E}f)(0) \prec 0$). En outre, pour tout $x \in E_3$,

$$\nabla^2(\mathbb{E}f)(x)\cdot(x^s, x^s) = -2||x^s||^4 < 0.$$

Ainsi, la condition (6) ne peut pas être vérifiée.

Nous venons de voir que $\mathbb{E}f$ n'a pas de mauvais point critique : les seuls points qui vérifient les conditions d'optimalité d'ordre 1 et 2 sont les solutions du problème de reconstruction de phase. Pour montrer que f elle-même n'a pas de mauvais point critique, une première idée serait de montrer qu'avec grande probabilité, pour tout x,

$$||\nabla f(x) - \nabla\mathbb{E}f(x)|| \text{ est très petite ;} \tag{7a}$$

$$|||\nabla^2 f(x) - \nabla^2\mathbb{E}f(x)||| \text{ est très petite.} \tag{7b}$$

pour une notion de « petitesse » à définir précisément, puis d'utiliser le fait que $\mathbb{E}f$ n'a pas de mauvais point critique pour en déduire que f non plus.

Telle quelle, cette approche ne fonctionne pas : de quelque manière qu'on définisse la « petitesse », les propriétés (7a) et (7b) ne sont, avec grande probabilité, pas vraies pour tous les x à la fois. Néanmoins, on peut montrer que ∇f et $\nabla^2 f$ partagent certaines propriétés de leurs espérances. Par exemple, le même genre d'arguments que dans la section 3 permet de montrer la propriété suivante.

Proposition 4.4. *Lorsque $m \geq Cn\log^3(n)$ pour une constante $C > 0$ assez grande, il est vrai, avec probabilité au moins $1 - O\left(\frac{1}{m}\right)$, que*

$$\nabla^2 f(x) \cdot (x^s, x^s) \leq -\alpha ||x^s||^4$$

pour tout $x \in Z_1 \overset{déf}{=} \{x \in \mathbb{C}^n, |\langle x, x^s\rangle| \leq \alpha ||x^s||^2, ||x|| \leq (1-\alpha)||x^s||\}$.

(Ici, $\alpha > 0$ est une constante dont on peut donner une valeur explicite.)

Cette propriété entraîne que f n'a pas de mauvais point critique sur l'ensemble Z_1. Si on définit d'autres ensembles $Z_2, Z_3, ...$ tels que l'union des Z_k recouvre $\mathbb{C}^n$ privé de l'ensemble des solutions et qu'on démontre sur $Z_2, Z_3, ...$ des propriétés similaires à celle que la proposition 4.4 énonce pour Z_1, on peut en déduire que f n'a pas de mauvais point critique du tout, ce qui conclut la démonstration.

4.3 Autres travaux

Cette technique de preuve ne s'applique bien sûr pas à tous les algorithmes : la non-existence de mauvais point critique est une propriété forte, que tous les algorithmes ne possèdent pas. En reconstruction de phase, l'algorithme de [Sun, Qu, and Wright, 2017a] que nous venons d'étudier est à ma connaissance le seul pour lequel ce phénomène ait été mis en évidence. Pour l'algorithme des projections alternées, par exemple, des expériences numériques suggèrent fortement qu'il existe presque toujours des mauvais points critiques[6]. Cela n'empêche d'ailleurs pas les projections alternées de fonctionner avec grande probabilité lorsque les vecteurs de mesure suivent une loi normale mais cela rend la technique de preuve que nous venons de voir inopérante.

En-dehors de la reconstruction de phase, on trouvera des exemples d'application de cette technique dans [Ge, Lee, and Ma, 2016], pour la complétion de matrices, dans [Bhojanapalli, Neyshabur, and Srebro, 2016], pour le *matrix sensing RIP*, dans [Sun, Qu, and Wright, 2017b], pour l'apprentissage de dictionnaire, et dans [Kawaguchi, 2016], pour les réseaux de neurones linéaires. Bien d'autres sont donnés dans l'article de revue [Zhang, Qu, and Wright, 2020]. Tous ces résultats, même s'ils partagent leur schéma de démonstration, sont très spécifiques au problème

6. sauf si $m \geq O(n^2)$ mais ce régime n'est pas très intéressant

considéré et à l'algorithme utilisé pour le résoudre. Quelques tentatives de généralisation ont été faites, notamment dans [Li and Tang, 2017] et [Ge, Jin, and Zheng, 2017], mais elles sont encore relativement rudimentaires.

Références

P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. SIAM Journal on Optimization, 16(2) :531–547, 2005.

S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In Advances in Neural Information Processing Systems 29, 2016.

E. J. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. Foundations of Computational Mathematics, 14(5) :1017–1026, 2014.

E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow : theory and algorithms. IEEE Transactions of Information Theory, 61(4) :1985–2007, 2015.

R. Chandra, Z. Zhong, J. Hontz, V. McCulloch, C. Studer, and T. Goldstein. Phasepack : A phase retrieval library. Asilomar Conference on Signals, Systems, and Computers, 2017.

Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. Communications on Pure and Applied Mathematics, 70(5) :2133–2150, 2017.

Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent : General statistical and algorithmic guarantees. preprint, 2015. https ://arxiv.org/abs/1509.03025.

R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems, pages 2973–2981. Curran Associates, Inc., 2016.

R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems : A unified geometric analysis. preprint, 2017. https ://arxiv.org/abs/1704.00708.

R. Gerchberg and W. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik, 35(2) :237–246, 1972.

P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In Symposium on the Theory of Computing, pages 665–674, 2013.

K. Kawaguchi. Deep learning without poor local minima. In Advances in neural information processing systems, pages 586–594, 2016.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. IEEE transactions on information theory, 56(6) :2980–2998, 2010.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent converges to minimizers. In Proceedings of the Conference on Computational Learning Theory, 2016.

Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. In 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pages 1235–1239, 2017.

M. Mondelli and A. Montanari. Fundamental limits of weak recovery with applications to phase retrieval. preprint, 2017. https ://arxiv.org/abs/1708.05932.

P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In Advances in Neural Information Processing Systems 26, pages 1796–2804, 2013.

I. Panageas and G. Piliouras. Gradient descent only converges to minimizers : Non-isolated critical points and invariant regions. preprint, 2016. https ://arxiv.org/abs/1605.00405.

J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. Foundations of Computational Mathematics, 2017a.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I : Overview and geometric picture. IEEE Transactions on Information Theory, 63(2), 2017b.

I. Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. IEEE Transactions on Information Theory, 64(5) :3301–3312, 2018.

G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving random systems of quadratic equations via truncated generalized gradient flow. To appear in IEEE Transactions on Information Theory, 2017.

H. Zhang and Y. Liang. Reshaped Wirtinger flow for solving quadratic systems of equations. In Advances in Neural Information Processing Systems 29, 2016.

Y. Zhang, Q. Qu, and J. Wright. From symmetry to geometry : Tractable nonconvex problems. preprint, 2020. https ://arxiv.org/abs/2007.06753.

T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In Advances in Neural Information Processing Systems 28, pages 559–567. Curran Associates, Inc., 2015.

Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. preprint, 2016. https ://arxiv.org/abs/1605.07051.