Méthodes de Burer-Monteiro

Jeudi 19 novembre 2020

1 Introduction et définitions

Dans cette dernière séance, nous allons étudier une famille d'algorithmes non-convexes, les méthodes de Burer-Monteiro, qui s'appliquent à tous les problèmes de reconstruction de matrices de bas rang vérifiant certaines hypothèses assez générales. On peut voir ces algorithmes comme une déconvexification des méthodes convexes. Par rapport aux deux séances précédentes, où nous avons considéré des problèmes de reconstruction précis, sous des hypothèses spécifiques, les garanties de correction que nous verrons ici ont un champ d'application bien plus large.

Dans cette introduction, nous donnerons d'abord les hypothèses que nous ferons sur les problèmes de reconstruction de matrices de bas rang (sous-section 1.1). Nous définirons ensuite les méthodes de Burer-Monteiro (sous-section 1.2) et observerons de manière préliminaire leur fonctionnement à travers quelques expériences numériques (sous-section 1.3).

1.1 Problèmes considérés

On considère comme précédemment des problèmes consistant à reconstruire une matrice X^s de rang faible à partir d'informations simples modélisées par l'appartenance à un certain ensemble $\mathcal{E}$:

$$\text{minimiser } \operatorname{rang}(X) \text{ pour } X \in \mathcal{E}. \tag{Min-rang}$$

Les algorithmes que nous allons étudier s'appliquent à ceux de ces problèmes qui admettent une relaxation convexe dotée de certaines propriétés. Rappelons que dans la première séance, nous avons vu qu'on peut approximer les problèmes de reconstruction de matrices de bas rang, non-convexes, par des problèmes convexes (les *relaxations convexes*). Assez souvent, la solution de la relaxation convexe est la même que celle du problème initial, de sorte que résoudre cette relaxation suffit à reconstruire X^s. On dit alors que la reconstruction est *exacte*.

Hypothèse 1. *Le problème* (Min-rang) *admet une relaxation convexe exacte.*

Dans cette séance, nous nous intéresserons aux problèmes qui admettent une relaxation convexe de la forme

$$\text{minimiser } \mathrm{Tr}(CX) \text{ pour } X \in \mathcal{E}', \tag{1}$$

où $C \in \mathcal{S}_n(\mathbb{R})$ est une matrice dépendant du problème considéré, qu'on appelle *matrice de coût*, et $\mathcal{E}'$ est un sous-ensemble de $\mathcal{S}_n^+(\mathbb{R})$, pour un certain entier $n \in \mathbb{N}^*$ (avec $\mathcal{S}_n^+(\mathbb{R})$ l'ensemble des matrices symétriques semi-définies positives de taille $n \times n$).

Nous supposerons en outre que l'ensemble $\mathcal{E}'$ satisfait la propriété suivante.

Hypothèse 2. *L'ensemble $\mathcal{E}'$ est compact. De plus, il est l'intersection de $\mathcal{S}_n^+(\mathbb{R})$ et d'un espace affine de dimension $m \geq 1$, de sorte que le problème (1) s'écrit*

$$\begin{aligned} &\textit{minimiser } \mathrm{Tr}(CX), \\ &\textit{avec } \mathcal{A}(X) = b, \\ &X \succeq 0, \end{aligned} \tag{SDP}$$

pour une certaine application linéaire $\mathcal{A} : \mathcal{S}_n(\mathbb{R}) \to \mathbb{R}^m$ et un certain vecteur $b \in \mathbb{R}^m$.

L'abréviation (SDP) signifie *SemiDefinite Program* (en français *problème d'optimisation semi-définie*).

Notons qu'on impose ici aux problèmes de minimisation de porter sur des matrices à coefficients réels et non complexes. Cela vise seulement à simplifier les énoncés : à de petites modifications près, les résultats que nous allons voir restent vrais pour des coefficients complexes.

Donnons quelques exemples de problèmes vérifiant les hypothèses 1 et 2.

1. Les problèmes de complétion de matrice admettent la relaxation

$$\begin{aligned} &\text{minimiser } ||X||_* \\ &\text{avec } X_{k,l} = X^s_{k,l}, \quad \forall (k,l) \in \Omega. \end{aligned}$$

   Ce problème peut se reformuler sous une forme (SDP) à cause de l'égalité suivante, valable pour toute $X \in \mathrm{Mat}(n_1, n_2)$:

$$||X||_* = \min \left\{ \frac{\mathrm{Tr}(Y) + \mathrm{Tr}(Z)}{2}, Y \in \mathrm{Mat}(n_1, n_1), Z \in \mathrm{Mat}(n_2, n_2), \begin{pmatrix} Y & X \\ {}^tX & Z \end{pmatrix} \succeq 0 \right\}.$$

   De plus, nous avons vu à la première séance que la relaxation est exacte avec grande probabilité sous des hypothèses adéquates.

2. Les problèmes de reconstruction de phase admettent plusieurs relaxations convexes de la forme (SDP) :

PhaseLift	*PhaseCut*
minimiser $\mathrm{Tr}(X)$,	minimiser $\mathrm{Tr}(MU)$,
avec $\langle X, v_k v_k^* \rangle = \lvert \langle x^s, v_k \rangle \rvert^2, \forall k \leq m,$	avec $U_{k,k} = 1, \forall k \leq m,$
$X \succeq 0.$	$U \succeq 0.$

Nous avons vu que *PhaseLift* était exacte avec grande probabilité sous certaines hypothèses. C'est également le cas de *PhaseCut*.

3. Le problème de synchronisation de phases admet une relaxation convexe de la forme

$$\begin{aligned} \text{minimiser } & -\mathrm{Tr}(CU), \\ \text{avec } & U_{k,k} = 1, \forall k \leq m, \\ & U \succeq 0. \end{aligned}$$

Dans le cas où les mesures sont contaminées par un bruit additif gaussien (comme à la séance précédente), cette relaxation est exacte avec grande probabilité si le niveau de bruit satisfait [1] $\sigma \leq c\sqrt{\frac{n}{\log(n)}}$ pour une certaine constante $c > 0$ [Zhong and Boumal, 2018].

1.2 Définition des méthodes de Burer-Monteiro

Notons $\mathcal{E}_{SDP}$ l'ensemble précédemment appelé $\mathcal{E}'$:

$$\mathcal{E}_{SDP} = \{X \in \mathcal{S}_n(\mathbb{R}), \mathcal{A}(X) = b, X \succeq 0\},$$

de sorte que le problème (SDP) peut se récrire comme

$$\text{minimiser } \mathrm{Tr}(CX) \text{ avec } X \in \mathcal{E}_{SDP}. \tag{SDP}$$

De nombreux algorithmes généraux existent pour résoudre les problèmes de cette forme. Nous en avons parlé à la première séance. Malheureusement, ces méthodes sont trop lentes pour être appliquées en grande dimension.

Les *méthodes de Burer-Monteiro* tentent de contourner cet obstacle [2] en déconvexifiant le problème (SDP). C'est à première vue contre-intuitif : si le problème (SDP) a été obtenu en « convexifiant » un problème de reconstruction de matrice de bas rang, ne risque-t-on pas, en

1. Notons que cette hypothèse est la même que celle sous laquelle nous avons démontré, à la séance précédente, la correction de la méthode non-convexe des puissances généralisées. Ce n'est pas un hasard : la correction de la méthode non-convexe est un élément clé pour la démonstration de l'exactitude de la relaxation convexe.

2. Rappelons que ce ne sont pas les seules [Ding, Yurtsever, Cevher, Tropp, and Udell, 2019; Yurtsever, Tropp, Fercoq, Udell, and Cevher, 2019].

déconvexifiant, de revenir à une formulation essentiellement identique au problème initial ? En fait, non. On aboutit à une famille de formulations, paramétrée par un entier p, dont certains éléments peuvent être plus simples à résoudre que la formulation initiale.

La déconvexification repose sur le fait que, sous l'hypothèse d'exactitude, la solution, X^s, du problème (SDP) est de rang faible. Pour tout $p \in \mathbb{N}^*$, notons

$$\mathcal{E}_{SDP,p} = \mathcal{E}_{SDP} \cap \{X \in \mathcal{S}_n(\mathbb{R}), \operatorname{rang}(X) \leq p\}.$$

Pour tout $p \geq \operatorname{rang}(X^s)$, le problème (SDP) est équivalent à

$$\text{minimiser } \operatorname{Tr}(CX) \text{ avec } X \in \mathcal{E}_{SDP,p}.$$

Informellement, l'ensemble $\mathcal{E}_{SDP,p}$ est de dimension beaucoup plus petite que $\mathcal{E}_{SDP,p}$ (si $p \ll n$ tout du moins) ; on peut donc espérer résoudre plus rapidement le problème de minimisation sur $\mathcal{E}_{SDP,p}$ que celui sur $\mathcal{E}_{SDP}$.

Pour résoudre le problème sur $\mathcal{E}_{SDP,p}$, le plus simple est de considérer la paramétrisation

$$V \in \mathcal{M}_p \to V^tV \in \mathcal{E}_{SDP,p},$$

avec

$$\mathcal{M}_p = \{V \in \operatorname{Mat}(n,p), \mathcal{A}(V^tV) = b\}.$$

Cette paramétrisation est bien surjective vers $\mathcal{E}_{SDP,p}$ car toute matrice $X \succeq 0$ de rang p peut s'écrire sous la forme $X = V^tV$ avec $V \in \operatorname{Mat}(n,p)$; de plus, si $\mathcal{A}(X) = b$, on a aussi $\mathcal{A}(V^tV) = b$ et donc $V \in \mathcal{M}_p$.

Cela permet de récrire le problème (SDP) comme

$$\text{minimiser } f_C(V) \overset{\text{déf}}{=} \operatorname{Tr}(CV^tV) \text{ avec } V \in \mathcal{M}_p. \qquad \text{(SDP factorisé)}$$

Dans ce dernier problème, l'inconnue V a seulement np coefficients, c'est-à-dire, si $p \ll n$, beaucoup moins que les n^2 coefficients de X. La manipuler est bien moins coûteux, en termes de temps de calcul, que manipuler X. En revanche, par rapport à (SDP), le problème (SDP factorisé) a le défaut de ne plus être convexe.

Même si leur succès n'est pas a priori certain à cause de la présence potentielle de points critiques, divers algorithmes simples peuvent être utilisés pour résoudre le problème (SDP factorisé). Nous appellerons *méthode de Burer-Monteiro* toute stratégie consistant à résoudre le problème (Min-rang) en passant par le problème (SDP factorisé) et en résolvant celui-ci à l'aide de n'importe lequel de ces algorithmes simples.

Dans ce cours, nous nous limiterons aux méthodes de Burer-Monteiro qui résolvent le problème (SDP factorisé) par *optimisation riemannienne*. Cela suppose que l'ensemble $\mathcal{M}_p$ soit une sous-variété riemannienne de $\operatorname{Mat}(n,p)$ [3]. En fait, nous ferons une hypothèse un peu plus forte.

3. Un-e lecteur/trice qui n'aurait pas de connaissances en géométrie riemannienne peut s'imaginer une sous-variété riemannienne comme une courbe ou une surface régulière dans $\operatorname{Mat}(n,p)$.

Hypothèse 3. *Pour tout $V \in \mathcal{M}_p$, l'application*

$$\tilde{\mathcal{A}} : V \in \mathrm{Mat}(n,p) \to \mathcal{A}(V^tV) \in \mathbb{R}^m$$

est de différentielle surjective.

Proposition 1.1. *Lorsque l'hypothèse 3 est vérifiée, $\mathcal{M}_p$ est une sous-variété riemannienne de $\mathrm{Mat}(n,p)$, de dimension $np - m$.*

Cette hypothèse est vérifiée notamment par les problèmes de reconstruction de phase (pour la relaxation convexe *PhaseCut*) et de synchronisation de phases. Pour la complétion de matrice, c'est moins clair ; cela semble dépendre de X^s.

Les algorithmes d'optimisation riemannienne sont en général des méthodes d'optimisation locale, qui partent d'un point de la variété considérée et le « déplacent » progressivement en utilisant les informations fournies par le gradient et, éventuellement, la hessienne de la fonction à minimiser. Ils dérivent la plupart du temps d'un algorithme d'optimisation itératif « standard » sur $\mathbb{R}^d$. Parmi les plus connus des algorithmes riemanniens, on peut citer la descente de gradient riemannienne (qui dérive de la descente de gradient sur $\mathbb{R}^d$) et la méthode des régions de confiance riemannienne (qui dérive de la méthode des régions de confiance sur $\mathbb{R}^d$). De nombreux autres existent [Absil, Mahony, and Sepulchre, 2009]. Chacun peut être plus ou moins adapté à une application donnée. Ici, nous ne ferons pas d'hypothèse spécifique sur l'algorithme choisi.

1.3 Résultats numériques

La question que nous allons discuter dans cette séance est :

Quand peut-on garantir que les méthodes de Burer-Monteiro fonctionnent ?

Nous tâcherons d'obtenir des garanties ne nécessitant essentiellement pas d'hypothèse sur $C, \mathcal{A}, b$ (outre les hypothèses 1, 2 et 3), de façon à ce que nos garanties s'appliquent au plus grand nombre de problèmes possible.

Nous devrons en revanche faire une hypothèse sur le paramètre p. En effet, la présence ou l'absence dans le problème (SDP factorisé) de points critiques susceptibles de nuire au bon fonctionnement des algorithmes riemanniens dépend fortement du choix de p. Idéalement, on aimerait que cette hypothèse soit vérifiée par des p aussi petits que possible : la dimension de l'inconnue V du problème (SDP factorisé) est proportionnelle à p et c'est donc pour les petites valeurs de p que les méthodes de Burer-Monteiro sont les plus rapides.

Commençons par nous faire une idée du fonctionnement des méthodes de Burer-Monteiro en regardant numériquement comment fonctionne l'une d'elles sur des problèmes de reconstruction de phase. Cette méthode consiste à utiliser la relaxation *PhaseCut* du problème de reconstruction

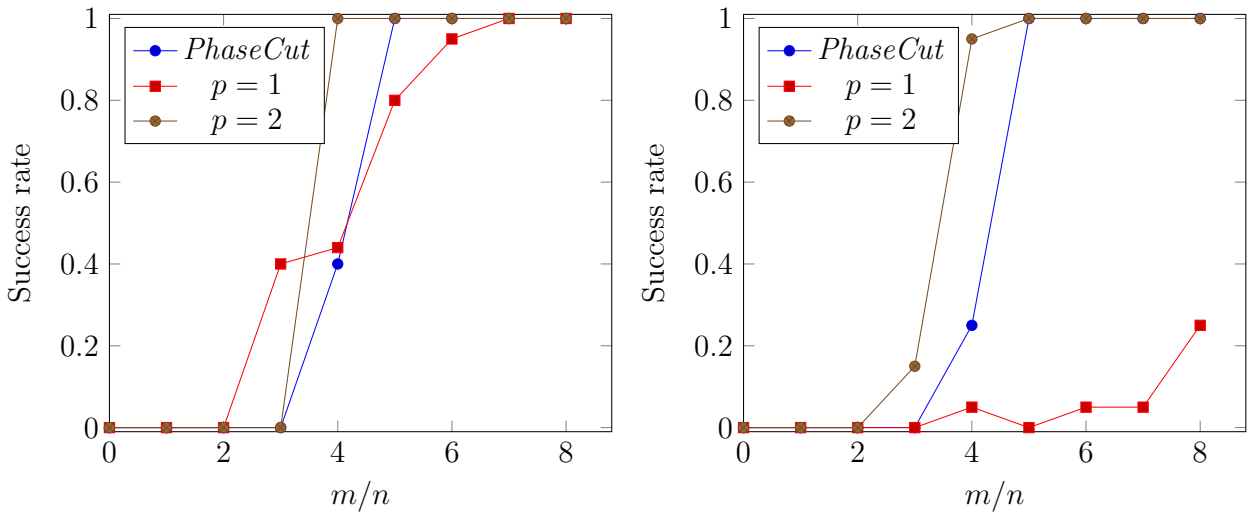


FIGURE 1 – Probabilité de succès en fonction de m/n de *PhaseCut*, d'une méthode de Burer-Monteiro avec $p = 1$, d'une méthode de Burer-Monteiro avec $p = 2$. La dimension du signal à reconstruire est $n = 32$. La figure de gauche correspond à des vecteurs de mesure $v_1, \ldots, v_m$ choisis indépendamment selon une loi normale et celle de droite à des vecteurs de mesure décrivant une « transformée en ondelettes ».

de phase et à résoudre le problème factorisé associé (SDP factorisé) au moyen d'une descente de gradient riemannienne. Ici, la solution cherchée est de rang 1 ; p peut donc a priori prendre n'importe quelle valeur entière supérieure ou égale à 1. La figure 1 représente les performances de cette méthode de Burer-Monteiro pour $p = 1$ et $p = 2$, appliquée en dimension $n = 32$, pour deux choix différents de vecteurs de mesure. À titre de comparaison est aussi représentée la performance de la méthode convexe *PhaseCut* (qui consiste à résoudre le problème (SDP) directement, sans le factoriser en (SDP factorisé)).

On observe que le fonctionnement de cette méthode de Burer-Monteiro dépend du choix des vecteurs de mesure. Avec $p = 1$, elle fonctionne mal dans l'un des deux cas considérés. Avec $p = 2$, elle fonctionne aussi bien que la méthode convexe dans chacun des deux cas.

D'autres expériences numériques ont été menées pour d'autres méthodes de Burer-Monteiro, appliquées à d'autres problèmes que la reconstruction de phase [Burer and Monteiro, 2003; Journée, Bach, Absil, and Sepulchre, 2010; Boumal, 2015]. Dans l'ensemble, elles aboutissent à des constatations similaires : pour les problèmes considérés, les méthodes fonctionnent très bien dès lors que p est « un peu plus grand » que le rang de la solution cherchée.

Ce comportement empirique favorable suggère qu'il n'est pas vain de chercher à établir des garanties de correction pour les méthodes de Burer-Monteiro. C'est l'objectif de la section 2.

Dans le section 3, nous discuterons de l'écart entre les garanties obtenues et le comportement empirique.

2 Garanties de succès lorsque $\frac{p(p+1)}{2} > m$

Dans cette section, nous allons expliquer les garanties de correction données dans [Boumal, Voroninski, and Bandeira, 2020], qui peuvent s'énoncer informellement comme suit : sous les hypothèses 1, 2, 3, si

$$\frac{p(p+1)}{2} > m,$$

alors les algorithmes riemanniens résolvent correctement presque tous les problèmes de la forme (SDP factorisé).

L'objet de la première sous-section qui suit est d'énoncer précisément ces garanties. La deuxième sous-section les démontre.

2.1 Énoncé précis

Dans la dernière partie de la deuxième séance, nous avons vu que lorsqu'on appliquait à une fonction régulière $f : \mathbb{R}^d \to \mathbb{R}$ une descente de gradient, on obtenait une suite convergente d'itérées (sous certaines hypothèses et sauf pour un ensemble de points initiaux de mesure nulle), dont la limite x_* était telle que

$$\nabla f(x_*) = 0 \qquad \text{et} \qquad \nabla^2 f(x_*) \succeq 0.$$

Cette propriété est vraie aussi pour au moins une partie des algorithmes riemanniens [Boumal, Absil, and Cartis, 2016; Criscitiello and Boumal, 2019] : appliqués au problème (SDP factorisé), ils permettent de trouver une matrice V_* telle que

$$\nabla_{\mathcal{M}_p} f_C(V_*) = 0 \qquad \text{et} \qquad \nabla^2_{\mathcal{M}_p} f_C(V_*) \succeq 0.$$

Ici, $\nabla_{\mathcal{M}_p}$ et $\nabla^2_{\mathcal{M}_p}$ sont le gradient et la hessienne de f_C restreinte à la variété $\mathcal{M}_p$. On appelle les $V_* \in \mathcal{M}_p$ vérifiant ces deux propriétés des *points critiques d'ordre* 2.

Les solutions [4] du problème (SDP factorisé) sont des point critique d'ordre 2 de f_C (c'est une propriété général des minimiseurs d'une fonction). S'il s'agit des *seuls* points critiques d'ordre 2, les méthodes de Burer-Monteiro sont assurées de résoudre correctement le problème (SDP factorisé) et donc le problème de départ (Min-rang).

4. J'utilise le pluriel car la solution n'est jamais unique : si V_* minimise f_C, alors, pour toute matrice unitaire $G \in \mathbb{R}^{p\times p}$, V_*G la minimise également.

Théorème 2.1 ([Boumal, Voroninski, and Bandeira, 2020]). *Soient $\mathcal{A} : \mathcal{S}_n(\mathbb{R}) \to \mathbb{R}^m, b \in \mathbb{R}^m$ fixés. On suppose que les hypothèses 1, 2 et 3 sont vraies et que*

$$\frac{p(p+1)}{2} > m.$$

Alors, pour toutes les matrices $C \in \mathcal{S}_n(\mathbb{R})$ hors d'un ensemble de mesure de Lebesgue nulle, le problème (SDP factorisé) n'a pas de point critique d'ordre 2 autre que les minimiseurs de f_C.

2.2 Démonstration

On suppose $\mathcal{A}, b$ fixés et les hypothèses 1, 2 et 3 vérifiées. Le résultat provient de la combinaison de deux lemmes.

Lemme 2.2. *Pour toute $V_* \in \mathcal{M}_p$, quelle que soit C, si*

$$\nabla_{\mathcal{M}_p} f_C(V_*) = 0, \quad \nabla^2_{\mathcal{M}_p} f_C(V_*) \succeq 0 \quad \textit{et} \quad \operatorname{rang}(V_*) < p,$$

alors V_* est solution de (SDP factorisé)

Lemme 2.3. *Supposons $\frac{p(p+1)}{2} > m$.*

Pour toutes les matrices $C \in \mathcal{S}_n(\mathbb{R})$ hors d'un ensemble de mesure de Lebesgue nulle, il n'existe pas $V_* \in \mathcal{M}_p$ telle que

$$\nabla_{\mathcal{M}_p} f_C(V_*) = 0, \quad \operatorname{rang}(V_*) = p.$$

En effet, si ces deux lemmes sont vrais, on obtient que, pour toutes les matrices C hors d'un ensemble de mesure nulle :

- le problème (SDP factorisé) n'a pas de point critique d'ordre 2 (ni même d'ordre 1) de rang p (lemme 2.3) ;
- le problème (SDP factorisé) n'a pas de point critique d'ordre 2 de rang inférieur à p autre que la solution (lemme 2.2).

Le problème (SDP factorisé) n'a donc pas de point critique autre que la solution (remarquons que les matrices de $\mathcal{M}_p$ sont de rang au plus p, puisqu'elles n'ont que p colonnes).

On peut donner du lemme 2.2 une interprétation géométrique en confondant (à tort) les notions de point critique d'ordre 2 et de minimiseur local. Si V_* est un minimiseur local de f_C sur $\mathcal{M}_p$, ${V_*}^t V_*$ est un minimiseur local de $X \to \langle C, X \rangle$ sur $\mathcal{E}_{SDP,p}$. La raison pour laquelle le lemme est vrai est que, si $\operatorname{rang}({V_*}^t V_*) < p$, alors l'ensemble des « directions de déplacement » possibles dans $\mathcal{E}_{SDP,p}$ forme un cône dont l'enveloppe convexe contient toutes les matrices satisfaisant les contraintes du problème (SDP). Cette propriété n'est pas évidente mais, si on l'admet, le lemme est établi : si ${V_*}^t V_*$ est un minimiseur local de $X \to \langle C, X \rangle$ sur $\mathcal{E}_{SDP,p}$, on doit avoir que

$\langle C, X - V_*{}^t V_* \rangle \geq 0$ pour tout X dans ce cône et donc $\langle C, X - V_*{}^t V_* \rangle \geq 0$ pour toute matrice X satisfaisant les contraintes du problème (SDP), de sorte que $V_*{}^t V_*$ est un minimiseur global de (SDP). Ainsi, V_* est bien solution du problème (SDP factorisé). Nous ne démontrerons pas ce lemme.

2.2.1 Idée de démonstration du lemme 2.3

Commençons par trouver une description explicite de l'ensemble des matrices de coût C pour lesquelles il existe $V_* \in \mathcal{M}_p$ telle que

$$\nabla_{\mathcal{M}_p} f_C(V_*) = 0 \quad \text{and} \quad \text{rang}(V_*) = p.$$

Pour toute $V_* \in \mathcal{M}_p$ de rang plein, l'application $V \in \mathcal{M}_p \to V^t V \in \mathcal{E}_{SDP,p}$ réalise essentiellement [5] un difféomorphisme au voisinage de V_* car sa différentielle est de rang localement constant. Le fait que le gradient de f_C sur $\mathcal{M}_p$ soit nul en V_* est alors équivalent au fait que le gradient de $X \to \langle C, X \rangle$ soit nul sur $\mathcal{E}_{SDP,p}$ en $V_*{}^t V_*$. Cela revient à dire que C est orthogonale à l'espace tangent à $\mathcal{E}_{SDP,p}$ en $V_*{}^t V_*$.

Ainsi, l'ensemble des matrices de coût C pour lesquelles il existe V_* telle que $\nabla_{\mathcal{M}_p} f_C(V_*) = 0$ et $\text{rang}(V_*) = p$ est

$$\bigcup_{M \in \mathcal{E}_{SDP,p}, \text{rang}(M)=p} (T_M \mathcal{E}_{SDP,p})^\perp . \tag{2}$$

(Pour toute $M \in \mathcal{E}_{SDP,p}$ de rang p, $T_M \mathcal{E}_{SDP,p}$ désigne l'espace tangent à $\mathcal{E}_{SDP,p}$ en M.)

Si on note N la dimension de $\mathcal{S}_n(\mathbb{R})$ et D celle de la variété $\{M \in \mathcal{E}_{SDP,p}, \text{rang}(M) = p\}$, $(T_M \mathcal{E}_{SDP,p})^\perp$ est de dimension $N - D$ pour toute M. L'ensemble précédent est donc une union d'espaces de dimension $N - D$, l'union étant paramétrée par une variété de dimension D. C'est donc intuitivement et informellement, comme l'illustre la figure 2a un ensemble de « dimension » [6]

$$(N - D) + D = N.$$

Un sous-ensemble de dimension N d'un espace vectoriel de dimension N n'est en général pas de mesure nulle. Comment peut-on améliorer le calcul précédent ?

Proposition 2.4. *Supposons que $\frac{p(p+1)}{2} > m$. Pour toute $M \in \mathcal{E}_{SDP,p}$ de rang p, il existe un segment inclus dans $\mathcal{E}_{SDP,p}$ dont l'intérieur contient M, sur lequel l'espace tangent à $\mathcal{E}_{SDP,p}$ est constant.*

5. L'application étant invariante par multiplication à droite par une matrice orthogonale, il faut pour obtenir exactement un difféomorphisme quotienter $\mathcal{M}_p$ par l'ensemble des matrices orthogonales de taille $p \times p$.

6. J'utilise des guillemets car l'ensemble en question n'est pas une variété et n'a donc pas de dimension, au sens usuel.

Cette proposition entraîne que, dans l'équation (2), la paramétrisation est redondante. L'ensemble peut en fait s'écrire comme une union d'espaces vectoriels de dimension $N - D$, paramétrée par un ensemble de dimension $D - 1$. Cet ensemble est donc de « dimension » au plus

$$(N - D) + (D - 1) = N - 1$$

et il est de mesure nulle dans $\mathcal{S}_n(\mathbb{R})$.

Démonstration de la proposition 2.4. Fixons $M \in \mathcal{E}_{SDP,p}$ de rang p.

Rappelons que

$$\mathcal{E}_{SDP,p} = \{X \succeq 0\} \cap \{X, \mathcal{A}(X) = b\} \cap \{X, \text{rang}(X) \leq p\}.$$

Comme M est de rang p, il existe un voisinage de M dans $\mathcal{S}_n(\mathbb{R})$ sur lequel toutes les matrices sont de rang au moins p et, de plus, toutes les matrices de rang p sont semi-définies positives (car M l'est). Ainsi, au voisinage de M, $\mathcal{E}_{SDP,p}$ coïncide avec l'ensemble plus simple

$$\{X, \mathcal{A}(X) = b\} \cap \{X, \text{rang}(X) = p\}.$$

Cette dernière expression (combinée avec l'hypothèse 3) permet de montrer que

$$\begin{aligned} T_M\mathcal{E}_{SDP,p} &= \text{Ker}\,(\mathcal{A}) \cap T_M\{X, \text{rang}(X) = p\} \\ &= \text{Ker}\,(\mathcal{A}) \cap \{X \in \mathcal{S}_n(\mathbb{R}), \langle v, Xv\rangle = 0 \text{ pour tout } v \in \text{Im}(M)^\perp\}. \end{aligned}$$

Comme cette expression ne dépend que de l'image de M, il suffit pour établir la proposition de montrer qu'il existe un segment inclus dans $\mathcal{E}_{SDP,p}$ dont tous les éléments ont la même image.

L'ensemble $\{X \in \mathcal{S}_n(\mathbb{R}), \text{Im}(X) \subset \text{Im}(M)\}$ est de dimension

$$\frac{\dim(M)(\dim(M) + 1)}{2} = \frac{p(p+1)}{2}.$$

Puisque $\frac{p(p+1)}{2} > m$, cet ensemble contient nécessairement une matrice $H \neq 0$ telle que

$$\mathcal{A}(H) = 0.$$

Pour tout $t \in \mathbb{R}$ assez proche de 0,

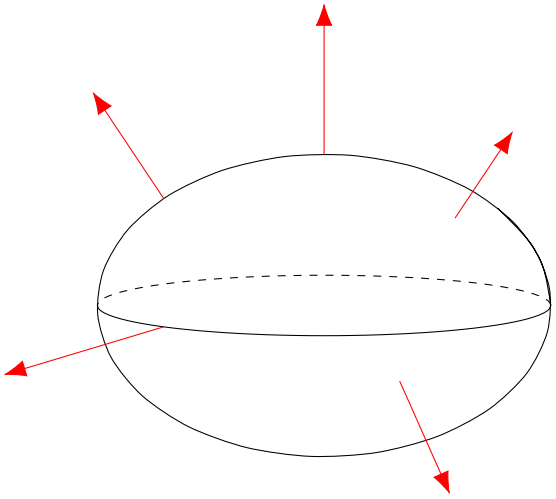
— $\text{rang}(M + tH) \geq \text{rang}(M) = p$ et $\text{Im}(M + tH) \subset \text{Im}(M)$ donc $\text{rang}(M + tH) = p$;

— $M + tH \succeq 0$;

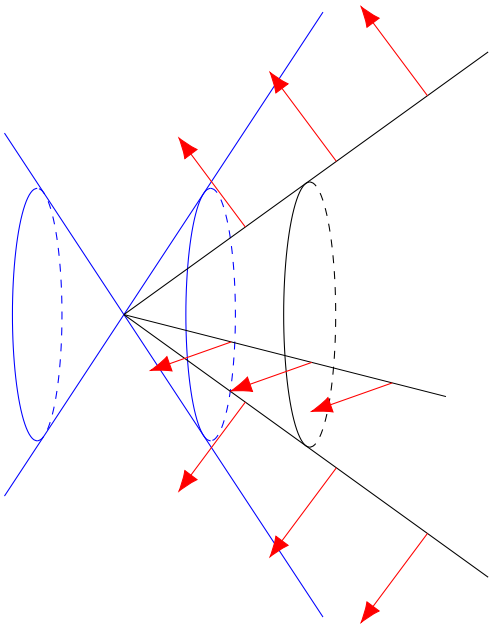
— $\mathcal{A}(M + tH) = \mathcal{A}(M) = b$.

Ainsi, pour $\epsilon > 0$ assez petit, $[M - \epsilon H; M + \epsilon H]$ est un segment inclus dans $\mathcal{E}_{SDP,p}$ le long duquel tous les éléments ont la même image que M (leur image est incluse dans $\text{Im}(M)$ et ils ont le même rang que M ; leur image est donc exactement égale à celle de M).

□



(a) Une variété de dimension 2 dans $\mathbb{R}^3$. L'union des directions orthogonales à l'un de ses espaces tangents est $\mathbb{R}^3$ tout entier.



(b) Une variété (ici un cône épointé, représenté en noir) dont chacun des points est à l'intérieur d'un segment le long duquel l'espace tangent est constant. L'union des directions orthogonales à l'un de ses espaces tangents est de dimension 2 (en l'occurrence le cône représenté en bleu).

3 Optimalité du résultat précédent

Le théorème que nous venons de démontrer affirme que les méthodes de Burer-Monteiro fonctionnent presque toujours lorsque $\frac{p(p+1)}{2} > m$. En ajoutant à C une petite perturbation aléatoire, on peut même obtenir des méthodes de Burer-Monteiro qui fonctionnent *toujours* au lieu de *presque toujours*, et même en temps polynomial en la précision souhaitée [Cifuentes and Moitra, 2019].

Malheureusement, la condition $\frac{p(p+1)}{2} > m$, qui ne peut être vérifiée que si $p \geq \sqrt{2m} + o(1)$, est assez décevante. En effet, l'équivalence entre le problème (SDP) et sa version factorisée (SDP factorisé) est vraie dès lors que $p \geq \operatorname{rang}(X^s)$ et on aurait donc voulu avoir également des garanties pour

$$\operatorname{rang}(X^s) \leq p < \sqrt{2m} + o(1).$$

Comme nous l'avons vu dans le paragraphe 1.3, les expériences numériques suggèrent que les méthodes de Burer-Monteiro peuvent bien fonctionner même si p est seulement légèrement

supérieur à $\text{rang}(X^s)$, ce qui est bien plus intéressant que choisir $p \geq \sqrt{2m}+o(1)$: dans beaucoup d'applications intéressantes, $\text{rang}(X^s) = O(1)$ tandis que $\sqrt{2m} = O(\sqrt{n})$.

Il importe donc de déterminer si la condition $\frac{p(p+1)}{2} > m$ du théorème 2.1 est optimale ou si elle peut être améliorée. Dans [Waldspurger and Waters, 2020], nous avons montré qu'elle était essentiellement optimale.

Théorème 3.1. *Supposons $\mathcal{A}, b$ fixés et les hypothèses 1, 2 et 3, ainsi que celle d'intersection minimale (voir plus bas), vérifiées.*

Notons $r_0 = \min\{\text{rang}(X), X \in \mathcal{E}_{SDP}\}$.

Pour tout p tel que

$$\frac{p(p+1)}{2} + pr_0 \leq m,$$

il existe un ensemble $\mathcal{C} \subset \mathcal{S}_n(\mathbb{R})$ de mesure de Lebesgue non-nulle tel que, pour toute $C \in \mathcal{C}$,

- $\text{rang}(X^s) = r_0$;
- *le problème* (SDP factorisé) *a un point critique d'ordre* 2 *qui n'est pas la solution cherchée.*

Pour se convaincre que ce théorème démontre bien la quasi-optimalité du théorème 2.1, il faut prendre en compte le fait que, pour les $\mathcal{A}, b$ qu'on rencontre en pratique, r_0 est généralement d'ordre 1 (exactement 1 pour la reconstruction de phase via *PhaseCut*) et que la condition

$$\frac{p(p+1)}{2} + pr_0 \leq m$$

est alors vérifiée pour tout $p \leq \sqrt{2m}+o(1)$, ce qui est essentiellement la condition complémentaire de celle du théorème 2.1.

Nous ne donnerons pas la définition précise de l'hypothèse d'« intersection minimale ». Elle requiert que l'intersection entre deux sous-espaces vectoriels particuliers de $\text{Mat}(n,p)$ soit la plus petite possible. Je n'ai pas d'interprétation géométrique satisfaisante de cette hypothèse mais elle est nécessaire pour la démonstration. Par chance, nous avons de bonnes raisons de penser qu'elle est satisfaite de manière « générique » (elle l'est, en tout cas, pour tous les exemples que nous avons étudiés), de sorte qu'elle ne constitue pas une restriction significative à l'applicabilité du théorème 3.1.

3.1 Idée de démonstration pour le théorème 3.1

La démonstration est en deux parties. Tout d'abord, on montre que, si une matrice C vérifie les propriétés de la conclusion du théorème et des conditions supplémentaires dites de *non-dégénérescence*, toutes les matrices dans un voisinage de C vérifient également ces propriétés. Ainsi, pour établir l'existence d'un ensemble de mesure non-nulle dont tous les éléments vérifient

ces propriétés, il suffit - c'est la deuxième partie - de montrer l'existence d'une seule matrice C vérifiant ces propriétés et les conditions de non-dégénérescence.

La première partie découle de propriétés élémentaires de géométrie différentielle. Donnons seulement une idée de la deuxième partie. Oublions les conditions de non-dégénérescence qui ne rajoutent pas de complexité majeure. Fixons $X_0 \in \mathcal{E}_{SDP}$ de rang r_0 et $V \in \mathcal{M}_p$ (essentiellement) quelconques ; nous allons voir qu'on peut trouver C telle que

1. $X^s = X_0$ et donc $\text{rang}(X^s) = r_0$;
2. V est un point critique d'ordre 2 non solution du problème (SDP factorisé).

Construire C nécessite de récrire ces propriétés en termes plus analytiques. La propriété 1 est impliquée par (et implique presque) le fait que C puisse se décomposer en

$$C = \mathcal{A}^*(g_1) + C_1,$$

pour un vecteur g_1 et une matrice C_1 telle que

$$C_1 X_0 = 0, \quad C_1 \succeq 0, \quad \text{rang}(C_1) = n - r_0.$$

Cela découle de la théorie classique de la dualité pour les problèmes d'optimisation semi-définie.

Pour le point 2, on observe que le gradient de f_C sur $\mathcal{M}_p$ en V est la projection sur $T_V\mathcal{M}_p$ du gradient de f_C sur $\text{Mat}(n,p)$ tout entier, qui vaut $2CV$. La nullité de ce gradient est donc équivalente à

$$CV \in (T_V\mathcal{M}_p)^\perp = \{\mathcal{A}^*(g)V, g \in \mathbb{R}^m\}.$$

Ainsi, le gradient est nul si et seulement si C s'écrit sous la forme

$$C = \mathcal{A}^*(g_2) + C_2$$

pour un vecteur g_2 et une matrice C_2 telle que $C_2 V = 0$.

On peut en outre montrer que, lorsque C s'écrit sous la forme précédente, la hessienne de f_C sur $\mathcal{M}_p$ en V vaut

$$\forall \dot{V} \in T_V\mathcal{M}_p, \quad \nabla^2 f_C(V) \cdot (\dot{V}, \dot{V}) = 2 \left\langle C_2, \dot{V}^t \dot{V} \right\rangle.$$

Ainsi, le fait que la hessienne soit semi-définie positive est équivalent à

$$\forall \dot{V} \in T_V\mathcal{M}_p, \quad \left\langle C_2, \dot{V}^t \dot{V} \right\rangle \geq 0.$$

Résumons : pour construire C, il suffit de trouver g_1, g_2, C_1, C_2 tels que

$$\mathcal{A}^*(g_1) + C_1 = \mathcal{A}^*(g_2) + C_2, \tag{3a}$$
$$C_1 X_0 = 0, \tag{3b}$$
$$C_1 \succeq 0, \tag{3c}$$
$$\operatorname{rang}(C_1) = n - r_0, \tag{3d}$$
$$C_2 V = 0, \tag{3e}$$
$$\forall \dot{V} \in T_V \mathcal{M}_p, \quad \left\langle C_2, \dot{V}^t \dot{V} \right\rangle \geq 0. \tag{3f}$$

En effet, si on les trouve, la matrice $C = \mathcal{A}^*(g_1) + C_1 = \mathcal{A}^*(g_2) + C_2$ vérifie les propriétés 1 et 2.

On peut sans perte de généralité imposer $g_2 = 0$. Ensuite, on remarque que, si on parvient à satisfaire les propriétés (3a) à (3e), on peut modifier C_1, C_2 en leur ajoutant une matrice semi-définie positive bien choisie de façon à ce que la propriété (3f) devienne vraie également. On peut donc se concentrer sur les propriétés (3a) à (3e).

Supposons pour simplifier que

$$\operatorname{Im}(X_0) = \{(x_1, \dots, x_{r_0}, 0, \dots, 0) \text{ avec } x_1, \dots, x_{r_0} \in \mathbb{R}\},$$
$$\operatorname{Im}(V) = \{(0, \dots, 0, x_{r_0+1}, \dots, x_{r_0+p}, 0, \dots, 0), \text{ avec } x_{r_0+1}, \dots, x_{r_0+p} \in \mathbb{R}\}.$$

Alors, pour g_1 fixé, il se trouve que l'existence de C_1, C_2 vérifiant les propriétés (3a) à (3e) est équivalente à ce que $\mathcal{A}^*(g_1)$ admette une décomposition en blocs de la forme

$$\mathcal{A}^*(g_1) = \begin{pmatrix} G_1 & 0 & G_2 \\ 0 & G_3 & G_4 \\ {}^tG_2 & {}^tG_4 & G_5 \end{pmatrix}, \tag{4}$$

avec $G_3 \in \mathcal{S}_p(\mathbb{R})$ définie négative. Or l'application

$$\begin{array}{rccccc} \Phi : & \mathbb{R}^m & \to & \mathbb{R}^{r_0 \times p} & \times & \mathcal{S}_p(\mathbb{R}) \\ & g & \to & \left((\mathcal{A}^*(g)_{k,l})_{\substack{1 \leq k \leq r_0 \\ r_0 < l \leq r_0+p}} \right. & , & \left. (\mathcal{A}^*(g)_{k,l})_{r_0 < k, l \leq r_0+p} \right) \end{array}$$

envoie un espace de dimension m vers un espace de dimension $\frac{p(p+1)}{2} + r_0 p \leq m$. Elle est donc a priori surjective (à coup sûr, en fait, grâce à l'hypothèse d'intersection minimale). Fixons $H \prec 0$ quelconque dans $\mathcal{S}_p(\mathbb{R})$. Choisissons pour g_1 un antécédent par Φ de $(0, H)$. La matrice $\mathcal{A}^*(g_1)$ est bien de la forme (4) (avec $G_3 = H$). Il existe donc C_1, C_2 de sorte que les propriétés (3a) à (3e) soient vérifiées, ce qui conclut.

4 Conclusion et questions ouvertes

Pour résumer, il semble qu'il y ait pour les méthodes de Burer-Monteiro trois régimes, aux frontières un peu floues :

- lorsque p est égal ou à peine supérieur à $\text{rang}(X^s)$, on observe qu'elles parviennent à résoudre certains problèmes mais on observe également des cas d'erreurs dans des situations intéressantes en pratique (Figure 1) ;
- lorsque p est supérieur à $\text{rang}(X^s)$ mais inférieur à $\sqrt{2m} + o(1)$, il semble numériquement qu'elles fonctionnent dans la quasi-totalité des situations concrètes dignes d'intérêt mais aucune garantie théorique générale ne l'explique ;
- lorsque p est supérieur à $\sqrt{2m}+o(1)$, on dispose de bonnes garanties théoriques (Théorème 2.1) mais, à ma connaissance, ces valeurs de p sont peu utilisées en pratique, puisqu'elles nécessitent de résoudre un problème de minimisation sur un espace d'inutilement grande dimension.

L'objectif de disposer d'algorithmes à la fois pratiques d'utilisation et pourvus de bonnes garanties théoriques n'est donc pas encore atteint. Les questions restant à résoudre sont, selon ma compréhension de ce sujet, principalement de deux types.

Il s'agit d'une part de questions plutôt géométriques, visant à comprendre le comportement des méthodes de Burer-Monteiro et l'existence de points critiques lorsque $\text{rang}(X^s) < p \leq \sqrt{2m}+o(1)$. Le théorème 3.1 affirme que, pour ces valeurs de p, si $\mathcal{A}$ et b fixés, il existe toujours des matrices de coût pour lesquelles des points critiques autres que la solution existent. De plus, ces matrices forment un ensemble de mesure non-nulle. Pourtant, on ne semble pas les rencontrer en pratique. Pourquoi ? Est-ce parce que l'ensemble est de volume infime, quoique strictement positif ? Est-ce parce que les méthodes de Burer-Monteiro parviennent à contourner les points critiques ? Dans ce cas, quelle technique utiliser pour le démontrer ?

L'autre axe concerne des aspects algorithmiques plus concrets. Peut-on relâcher les hypothèses 2 et 3, pour qu'elles s'appliquent à davantage de problèmes ? Pour l'hypothèse 2, cette question a déjà été abordée dans [Bhojanapalli, Boumal, Jain, and Netrapalli, 2018]. Indépendamment, quels algorithmes riemanniens sont les plus adaptés pour trouver approximativement un point critique d'ordre 2 ? Cette question est cruciale pour les applications mais semble délicate. Le possible mauvais conditionnement du problème (SDP factorisé) près de sa solution est notamment source de difficultés (voir par exemple [Tong, Ma, and Chi, 2020] et les références de cet article).

Références

P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. Princeton University Press, 2009.

S. Bhojanapalli, N. Boumal, P. Jain, and P. Netrapalli. Smoothed analysis for low-rank solutions to semidefinite programs in quadratic penalty form. In Proceedings of the 31st Conference On Learning Theory, pages 3243–3270, 2018.

N. Boumal. A riemannian low-rank method for optimization over semidefinite matrices with block-diagonal constraints. Technical report, Inria, 2015. http ://arxiv.org/abs/1506.00575.

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. IMA Journal of Numerical Analysis, 2016.

N. Boumal, V. Voroninski, and A. S. Bandeira. Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs. Communications on Pure and Applied Mathematics, 73(3) :581–608, 2020.

S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Mathematical Programming, 95(2) :329–357, 2003.

D. Cifuentes and A. Moitra. Polynomial time guarantees for the Burer-Monteiro method. preprint, 2019. https ://arxiv.org/abs/1912.01745.

C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In Advances in Neural Information Processing Systems, pages 5987–5997, 2019.

L. Ding, A. Yurtsever, V. Cevher, J. A. Tropp, and M. Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. preprint, 2019. https ://arxiv.org/abs/1902.03373.

M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. SIAM Journal on Optimization, 20(5) :2327–2351, 2010.

T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. preprint, 2020. https ://arxiv.org/abs/2005.08898.

I. Waldspurger and A. Waters. Rank optimality for the Burer-Monteiro factorization. To appear in SIAM journal on optimization, 2020.

A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, and V. Cevher. Scalable semidefinite programming. preprint, 2019. https ://arxiv.org/abs/1912.02949.

Y. Zhong and N. Boumal. Near-optimal bounds for phase synchronization. SIAM Journal on Optimization, 28(2) :989–1016, 2018.