Régularisation TV pour le consensus robuste dans les systèmes distribués

Walid BENAMEUR[1], Pascal BIANCHI[2], Jérémie JAKUBOWICZ[1]

[1]SAMOVAR, Telecom Sud-Paris & CNRS
9 rue Charles Fourier, 91011 Evry, France

[2]LTCI, Telecom ParisTech & CNRS
46 rue Barrault, 75013 Paris, France

walid.benameur@telecom-sudparis.fr, pascal.bianchi@telecom-paristech.fr, jeremie.jakubowicz@telecom-sudparis.fr

Résumé – Considérons un réseau d'agents dont l'objectif est de trouver un consensus sur le minimiseur d'une certaine fonction inconnue. Un cas particulier important est obtenu lorsque ce minimiseur est la moyenne de valeurs localement observées par les agents. On s'intéresse à des algorithmes distribués permettant d'estimer la solution par le biais d'échanges entre les agents, supposés connectés par un graphe. La littérature est riche en algorithmes distribués résolvant un tel problème. Toutefois, ces algorithmes s'avèrent très sensibles à de possibles comportements déficients de certains agents (agents malveillants ou agents "têtus" refusant d'actualiser leur estimée). Notre objectif est de proposer des algorithmes robustes. Nous introduisons une relaxation du problème initial fondée sur la variation totale. Nous démontrons que les solutions du problème initial et du problème relaxé coïncident, sous certaines conditions de régularité vérifiables. Nous fournissons deux algorithmes distribués conduisant à la solution. Enfin, nous testons la robustesse de nos algorithmes en présence d'un agent têtu injectant en permanence une estimée incohérente dans le réseau : nous montrons que, indépendamment de l'amplitude de l'estimée incohérente, les estimées sont maintenues dans un voisinage du consensus souhaité. Des résultats numériques confirment la robustesse de nos algorithmes de cet article.

Abstract – Consider a network of agents seeking consensus on a given function minimizer. As an important particular case, we mention the case where this minimizer is the average of all values held by the agents. We focus on distributed algorithms that achieve consensus using available links in the network. There is a huge literature on this problem. However, existing algorithms turn out fragile to ill-functioning agents (malicious agents or "stubborn agents" refusing to update). Our goal is to provide robust algorithms. We introduce a total variation based relaxation of the initial problem. And then we show that both initial problem and the relaxed version share the same solutions, provided some verifiable regularity conditions are fulfilled. We provided two distributed algorithms leading to the solution. Last, we challenge the proposed algorithms robustness by introducing a stubborn agent that keeps injecting an incoherent estimate in the network. We prove that estimates are kept in the vicinity of the sought consensus. Numerical results back up this claim.

1 Introduction

On considère un réseau modélisé par un graphe non-dirigé $G = (V, E)$ où V représente un ensemble fini d'agents et où $\{v, w\}$ appartient à E si et seulement si les agents v et w peuvent communiquer. Nous étudions le problème d'optimisation :

$$\min_{x\in\mathbb{R}} \sum_{v\in V} F_v(x) \tag{1}$$

où $F_v : \mathbb{R} \to \mathbb{R}$ est une fonction convexe coercive connue par l'agent v seulement, et qui représente le regret de l'agent v lorsque le réseau se trouve dans l'état x. Nous nous limitons dans cet article au cas où x est réel afin de ne pas compliquer les notations. Bien que nos résultats soient généraux, nous accordons une importance toute particulière à deux cas de figure pratiques : le cas du *consensus sur la moyenne (CMo)* et le cas du *consensus sur la médiane (CMe)*. Les fonctions F_v sont respectivement données par :

$$(CMo) : \; F_v(x) = \frac{1}{2}(x - x_0(v))^2 \, , \tag{2}$$

$$(CMe) : \; F_v(x) = |x - x_0(v)| \, , \tag{3}$$

où $x_0(v)$ représente une valeur initiale dont dispose l'agent v uniquement. Dans ce cas, le problème (1) revient respectivement à calculer la moyenne $\overline{x}_0 = (1/|V|)\sum_v x_0(v)$ (où $|V|$ est le cardinal de V) ou la médiane du vecteur x_0. Le problème du calcul distribué de la moyenne sur un réseau a fait l'objet de maints travaux [1, 2]. En comparaison, le calcul de la médiane a été moins étudié, mais s'avère être une perspective intéressante afin d'améliorer la robustesse aux valeurs aberrantes.

Chaque agent v génère une suite d'estimées $x_n(v)$. L'objectif de cet article est de proposer et d'analyser des algorithmes distribués qui entrainent les estimées vers un minimiseur commun de (1) quand $n \to \infty$. De tels algorithmes existent déjà dans la litérature [1, 3]. Toutefois, la plupart des algorithmes existants se révèlent très sensibles à la présence d'imperfections dans le réseau (agents malveillants ou agents "têtus" refusant d'actualiser leur valeur ou de coopérer). Nous recherchons ici des algorithmes robustes, c'est à dire que l'estimée finale des agents sains doit au moins rester dans un voisinage du consensus souhaité, y compris en présence d'un réseau imparfait.

2 Notations et résultats préliminaires

Soit $d(v)$ le degré d'un sommet v (son nombre de voisins). Nous aurons parfois besoin de donner une orientation arbitraire au graphe, nous noterons alors $(V, \vec{E})$ le graphe dirigé. Toutefois, notons bien que nos résultats ne dépendront pas de l'orientation choisie pour $\vec{E}$. A toute fonction $x \in \mathbb{R}^V$, on associe le gradient $\mathrm{grad}(x) \in \mathbb{R}^{\vec{E}}$ défini par $\mathrm{grad}(x)(v,w) = x(w) - x(v)$. Pour tout $A \subset V$, on définit 1_A comme la fonction indicatrice de A (égale à un sur A et zéro ailleurs). Désignons par $\mathbb{R}_0^V$ l'ensemble $\{x \in \mathbb{R}^V : \langle x, 1_V \rangle = 0\}$ des fonctions de moyenne nulle sur V. On vérifie sans peine que $\|x\|_{\mathrm{TV}} := \sum_{e \in \vec{E}} |\mathrm{grad}\, x|(e)$ est une semi-norme sur $\mathbb{R}^V$ et une norme sur $\mathbb{R}_0^V$ si G est connexe. On introduit la norme duale $\|u\|_* = \max_{\|x\|_{\mathrm{TV}} \leq 1} \langle x, u \rangle$ où $\langle x, u \rangle$ représente le produit scalaire usuel sur $\mathbb{R}_0^V$. Nous notons $B_* = \{u \in \mathbb{R}_0^V : \|u\|_* \leq 1\}$ la boule unité. On note $\mathrm{Per}(A) = \|1_A\|_{\mathrm{TV}}$ le périmètre d'un ensemble $A \subset V$ et $G(A)$ la restriction du graphe G aux seuls sommets de A. Les conditions de validité des résultats qui vont suivre mettent en jeu la norme duale. Afin de vérifier ces conditions, il est nécessaire de disposer d'un algorithme efficace de calcul de cette norme. Nous en donnons une expression possible ci-dessous.

Proposition 1 *Pour tout $u \in \mathbb{R}_0^V$,*

$$\|u\|_* = \max \left\{ \frac{|\langle u, 1_A \rangle|}{\mathrm{Per}(A)} : \emptyset \subsetneq A \subset V, |A| \leq \frac{|V|}{2}, G(A) \textit{ connexe} \right\}$$

Cette expresion montre que pour des graphes simples comme les chaînes ou les cycles, la norme duale se calcule par simple énumération de sous-ensembles connexes. Cette énumération peut devenir fastidieuse pour des graphes quelconques. Dans ce cas l'algorithme qui suit permet de calculer la norme en un nombre d'iterations qui ne peut dépasser $|E|$ sachant que chaque itération se ramène à un calcul d'une coupe minimum dans un graphe.

Algorithme 0 (calcul de la norme duale de $u \in \mathbb{R}_0^V$) :
- Choisir un ensemble $\emptyset \subsetneq A_0 \subsetneq V$, poser $\lambda_1 = \frac{|\langle u, 1_{A_0} \rangle|}{\mathrm{Per}(A_0)}$ et $i = 1$.
- Répéter
  - $A_i = \arg\max_{A \subset V} \langle u, 1_A \rangle - \lambda_i \,\mathrm{Per}(A)$
  - $\lambda_{i+1} = \frac{\langle u, 1_{A_i} \rangle}{\mathrm{Per}(A_i)}$ et $i = i + 1$
- Jusqu'à $\lambda_i = \lambda_{i-1}$
- Retourner λ_i.

3 Régularisation par variation totale

De façon très générale, notre approche consiste à remplacer le problème (1) par :

$$\min_{x \in \mathbb{R}^V} \sum_{v \in V} F_v(x(v)) + P(x) \tag{4}$$

où $P : \mathbb{R}^V \to \mathbb{R}$ est un terme de régularisation convexe pénalisant les fonctions $x \in \mathbb{R}^V$ qui s'écartent de l'espace de consensus $\mathcal{C} := \mathrm{Vect}(1_V)$. Dans le cas où l'on choisit pour P la fonction indicatrice de $\mathcal{C}$ (nulle sur $\mathcal{C}$ et égale à $+\infty$ sinon), les problèmes (1) et (4) sont rigoureusement équivalents. D'un point de vue algorithmique, un tel choix de P revient à imposer le consensus *coûte que coûte* (le prix d'un écart au consensus est infini). Toutefois, en présence d'agents irréguliers, il s'avère intéressant de relaxer le prix d'un écart au consensus, afin de laisser aux agents réguliers la possibilité d'être en désaccord avec les agents irréguliers. Une telle relaxation a naturellement un défaut : nous ne pouvons plus espérer que les minimiseurs de (1) et (4) coïncident pour tout choix de $(F_v)_{v \in V}$. Heureusement, nous montrons qu'en fixant $P(x) = \lambda \|x\|_{\mathrm{TV}}$, les minimiseurs de (1) et (4) coïncident *au moins pour une certaine classe de fonctions* $(F_v)_{v \in V}$ que nous caractérisons. Dans les scénarios (CMo) et (CMe) le problème devient respectivement :

$$(a) \quad : \quad \min_{x \in \mathbb{R}^V} \frac{1}{2}\|x - x_0\|_2^2 + \lambda \|x\|_{\mathrm{TV}} \tag{5}$$

$$(b) \quad : \quad \min_{x \in \mathbb{R}^V} \|x - x_0\|_1 + \lambda \|x\|_{\mathrm{TV}} \tag{6}$$

On fixe dorénavant $P(x) = \lambda \|x\|_{\mathrm{TV}}$. Notre objectif est triple : (i) déterminer les situations où les minimiseurs de (1) et (4) coïncident ; (ii) proposer des algorithmes distribués pour résoudre (4) ; (iii) tester la robustesse de ces algorithmes en présence d'agents irréguliers.

On définit pour tout $x \in \mathbb{R}^V$, $F(x) = \sum_v F_v(x(v))$. Pour tout $x^\star \in \mathbb{R}$, si ∂ représente le sous-différentiel,

$$\partial F(x^\star 1_V) = \left\{ u \in \mathbb{R}^V : \forall v \in V,\, u(v) \in \partial F_v(x^\star) \right\} .$$

Plus simplement, notons que si tous les F_v sont dérivables, $\partial F(x^\star 1_V)$ est réduit au singleton $\{(F_v'(x^\star))_{v \in V}\}$.

Théorème 1 *Parmi les assertions suivantes, 1) 2) 3) sont équivalentes et impliquent 4).*

1. *$x^\star 1_V$ est un minimiseur de (4) ;*
2. *$\partial F(x^\star 1_V) \cap \lambda B_*$ est non-vide ;*
3. *Il existe $u \in \partial F(x^\star 1_V)$ tel que $\sum_{v \in V} u(v) = 0$ et pour tout $A \subset V$, $\sum_{v \in A} u(v) \leq \lambda\, \mathrm{Per}(A)$;*
4. *$x^\star$ est un minimiseur de (1).*

Voyons maintenant les conséquences de ce résultat dans les cas particuliers (CMo) et (CMe).

Proposition 2 (CMo) *Soit $\overline{x}_0 = \sum_{v \in V} x_0(v)/|V|$. Les assertions suivantes sont équivalentes.*

1. *$\overline{x}_0 1_V$ est l'unique minimiseur de (6)(a) ;*
2. *$x_0 - \overline{x}_0\, 1_V \in \lambda B_*$;*
3. *Pour tout $A \subset V$, $\left|\frac{\sum_{v\in A} x_0(v)}{|A|} - \bar{x}_0\right| \le \lambda \frac{\mathrm{Per}(A)}{|A|}$.*

La proposition 2 fournit les conditions sous lesquelles le minimum du problème relaxé (6)(a) coïncide avec le consensus $\overline{x}_0$ souhaité. Le point 2) assure que les minimiseurs coïncident *exactement* dès lors que λ est choisi plus grand que $\|x_0 - \bar{x}_0\|_*$. Le point 3) indique jusqu'à quel point les moyennes *locales* de x_0 peuvent fluctuer autour de $\overline{x}_0$. Plus le rapport $\mathrm{Per}(A)/|A|$ est grand, plus on peut admettre que la moyenne locale de x_0 sur A soit loin de $\overline{x}_0$. Intuitivement, $\mathrm{Per}(A)/|A|$ mesure à quel point une région A est connectée à l'ensemble du réseau : $\mathrm{Per}(A)$ est le nombre d'arêtes connectant A à son complémentaire $V\backslash A$ et $|A|$ est une mesure de sa taille. Ainsi, dire que $\mathrm{Per}(A)/|A|$ est grand revient à dire que, relativement à sa taille, la région A admet beaucoup de connections avec l'extérieur. Dans ce cas, il sera "plus facile" pour cette région d'amener l'ensemble de ses nœuds vers le bon consensus global $\overline{x}_0$.

Désignons par $\mathcal{U}$ l'ensemble des vecteurs $u \in \mathbb{R}^V$ qui contiennent une moitié de 1 et une moitié de -1 (et un 0 dans le cas où $|V|$ est impair). On pose $\lambda_0 = \max\{\|u\|_* \,:\, u \in \mathcal{U}\}$.

Proposition 3 (CMe) *Pour tout $\lambda \ge \lambda_0$, l'ensemble des minimiseurs de (6)(b) est égal à* $\mathrm{mediane}(x_0)1_V$.

4 Algorithmes proposés

Notons que la fonction $P(x) = \lambda\|x\|_{\mathrm{TV}}$ *n'est pas différentiable* : une première approche consiste à écrire l'algorithme du *sous-gradient* associé au problème (4). Cet algorithme donne lieu de manière naturelle à une implémentation distribuée sur le graphe, où chaque nœud v maintient une estimée $x_n(v)$ du minimiseur et combine à chaque instant son estimée avec celle de ses voisins :

Algorithme 1 :

$$x_{n+1}(v) = x_n(v) + \gamma_n \Big[g_n(v) + \lambda \sum_{w\sim v} \mathrm{sign}\left(x_n(w) - x_n(v)\right)\Big]$$

où $g_n(v) \in \partial F_v(x_n(v))$, typiquement $g_n(v) = x_n(v) - x_0(v)$ dans le cas (CMo). On démontre par des arguments standards que la fonction x_n converge vers un minimiseur de (4) sous l'hypothèse de pas γ_n décroissants. L'algorithme du sous-gradient est réputé pour être assez peu efficient en terme de vitesse de convergence. Plusieurs alternatives existent afin d'accélérer la vitesse tout en conservant une implémentation distribuée. Nous proposons ici un algorithme qui peut être vu comme un cas particulier de l'*Alternating Direction of Multipliers Method* [4]. Appelons $\mathrm{proj}_{[-\omega,\omega]}(x)$ le projeté de x sur l'intervalle $[-\omega,\omega]$ et notons $\mathrm{prox}_{f,\rho}(x) = \arg\min_y f(y) + \frac{\rho}{2}(y-x)^2$ l'opérateur proximal associé à une fonction réelle f quelconque.

Algorithme 2 A chaque instant n, l'agent $v \in R$ reçoit $(x_n(w) : w \sim v)$. Pour $v \sim w$,

$$\mu_{n+1}(w,v) = \mathrm{proj}_{[-2\lambda/\rho, 2\lambda/\rho]}\left(\mu_n(w,v) + x_n(w) - x_n(v)\right) \quad (7)$$

$$x_{n+1}(v) = \mathrm{prox}_{F_v,\,\rho d(v)}\left(x_n(v) + \tfrac{3}{2}\,\tilde{\mu}_{n+1}(v) - \tfrac{1}{2}\,\tilde{\mu}_n(v)\right) \quad (8)$$

où l'on pose $\tilde{\mu}_n(v) = \frac{1}{d(v)} \sum_{w\sim v} \mu_n(w,v)$.

5 Test de robustesse

Nous prétendons que les algorithmes proposés sont plus robustes à la présence de comportements déficients de certains agents que la plupart des méthodes existant dans la littérature [1]. D'une part nous démontrons un résultat allant dans ce sens et d'autre part nous fournissons plus loin des expériences numériques illustrant cette propriété de robustesse des algorithmes proposés.

Scénario : *On considère le problème (CMo) du consensus sur la moyenne. Le graphe G est supposé complet. Soit $t \in V$ un agent arbitraire. Chaque agent $v \in V\backslash\{t\}$ actualise son estimée $x_n(v)$ selon l'algorithme 1 (resp. l'algorithme 2). L'agent t maintient quant à lui sa valeur initiale, soit $x_n(t) = x_0(t)$ pour tout n.*

Soit $x_{0,-t}$ la restriction de x_0 à $V\backslash\{t\}$ et

$$\overline{x}_{0,-t} = \sum_{v\neq t} x_0(v)/(|V|-1)$$

la moyenne des valeurs initiales des agents sains. Soit $\|\,.\,\|_{*,-t}$ la norme duale associée au graphe $G(V\backslash\{t\})$ et $B_{*,-t}$ la boule unité associée.

Théorème 2 *Supposons $(x_{0,-t} - \overline{x}_{0,-t} 1_{V\backslash\{t\}}) \in \lambda B_{*,-t}$. Pour tout $v \in V\backslash\{t\}$, $x_n(v)$ converge vers :*

$$\overline{x} = \begin{cases} x_0(t) & \text{si } |\overline{\boldsymbol{x}}^{0,R} - x_0(t)| \le \lambda \\ \overline{x}_{0,-t} + \lambda & \text{si } \overline{x}_{0,-t} < x_0(t) - \lambda \\ \overline{x}_{0,-t} - \lambda & \text{si } \overline{x}_{0,-t} > x_0(t) + \lambda\,. \end{cases}$$

L'hypothèse posée dans le théorème ci-dessus est une condition de régularité similaire à celle de la proposition 2 : en l'absence de l'agent têtu t, l'algorithme aurait conduit les agents vers le consensus souhaité $\overline{x}_{0,-t}$. En présence de l'agent têtu, les agents sains s'accordent toujours au consensus, mais celui-ci n'est plus $\overline{x}_{0,-t}$. En revanche, l'amplitude du biais $\overline{x} - \overline{x}_{0,-t}$ est *majorée par* λ. Ainsi, λ joue le rôle d'une marge maximale d'erreur. Diminuer λ réduit l'écart au consensus souhaité : le compromis à trouver provient du fait qu'en diminuant λ, on restreint la classe des fonctions $x_{0,-t}$ qui satisfont la condition de régularité.

On se propose maintenant de vérifier cette propriété de robustesse de façon numérique. Pour cela nous précisons quelques ordre de grandeurs. Ces expériences ont été menées dans un cadre plus exhaustif que celui présenté ici, suffisamment large cependant pour donner une bonne idée de la robustesse des algorithmes. Considérons un graphe à $N = 99$ sommets. L'ensemble V est assimilé à $\{1, \ldots, N\}$. On suppose que la structure des arêtes correspond au cas du graphe complet, c'est-à-dire que les $N(N-1)/2$ arêtes sont présentes, tous les agents étant connectés entre eux. On suppose dans un premier temps que tous les agents sont de bonne volonté et nous procédons suivant les itérations des algorithmes 1 et 2. On trace alors la courbe $n \to \log \|J^{\perp} x_n\|$ où $J^{\perp}$ correspond au projecteur orthogonal $I - J$ sur $1^{\perp}$, avec $J = 11^T/N$ projecteur orthogonal sur la droite engendrée par le vecteur 1. De sorte que lorsque la courbe $n \to \log \|J^{\perp} x_n\|$ prend des valeurs très négatives, cela implique que le consensus est quasiment atteint, on peut alors se référer à $\bar{x}_n$ pour avoir la valeur de ce consensus.

On considère le vecteur initial $x_0 = \sin(5\pi v/N)$ où v décrit V. Et on se donne $\lambda = 1$. On calcule la norme duale $\|x_0 - \bar{x}_0 1\|_*$ grâce à l'algorithme 0 et on obtient $\|x_0 - \bar{x}_0 1\|_* \approx 0.013$. On en déduit qu'on se trouve dans le cas $\lambda > \|x_0 - \bar{x}_0 1\|_*$ et que par conséquent, en vertu du Théorème 1, $\bar{x}_0 1$ est le minimiseur cherché pour le problème (CMo). On calcule $\bar{x}_0 \approx 0.1271$ et $\mathrm{mediane}\, x_0 \approx 0.2817$. On fixe le nombre d'itérations à 300.

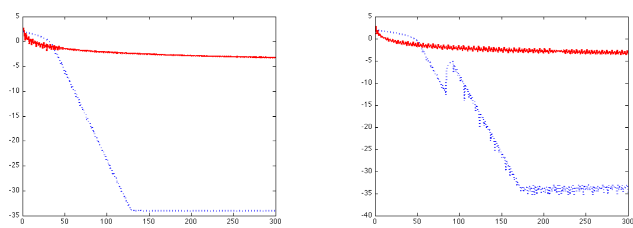

FIGURE 1 – Absence d'agent malveillant. Gauche : Le problème de (CMo). Droite : Le problème (CMe). Pour chaque problème on représente $n \to \log \|J^{\perp} x_n\|$ pour l'algorithme 1 (en rouge, trait plein) et 2 (en bleu, pointillés)

Il apparaît clairement sur la figure 2 que pour (CMo) comme pour (CMe), pour l'algorithme 1 comme pour l'algorithme 2, le consensus est atteint. Pour l'algorithme 2 on note une convergence vers le consensus exponentiellement rapide (dans la mesure où la courbe représentée correspond à un logarithme), jusqu'à ce que la précision machine soit atteinte. Pour l'algorithme 1 la convergence est beaucoup plus lente et présente des phénomènes d'oscillation, caractéristiques. On vérifie alors la valeur de $\bar{x}_n$ pour $n = 300$ et on trouve pour (CMo) par l'algorithme 1 et 2 $\bar{x}_n \approx 0.1271$. On pourrait en fait montrer facilement que la vraie moyenne est conservée le long des itérations. Pour le problème (CMe) on trouve $\bar{x}_n = .2809$ pour l'algorithme 1 et $\bar{x}_n = .2818$ pour l'algorithme 2. Tout est donc conforme à ce qu'on attendait des résultats précédents.

On introduit maintenant un agent têtu et on reprend les mêmes expériences. On suppose donc maintenant que l'agent 1 est têtu et qu'il se fige à la valeur 10 qui est d'un ordre de grandeur supérieure aux valeurs de x_0 des agents réguliers. Pourtant, on

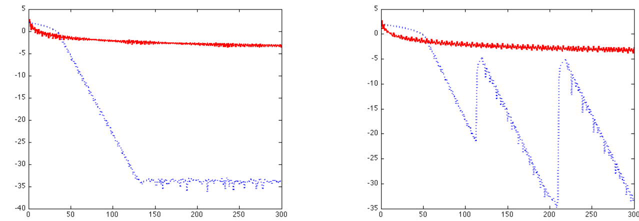

FIGURE 2 – Avec un agent malveillant. Gauche : Le problème de (CMo). Droite : Le problème (CMe). Pour chaque problème on représente $n \to \log \|J^{\perp} x_n\|$ pour l'algorithme 1 (en rouge, trait plein) et 2 (en bleu, pointillés)

note la convergence vers le consensus et on vérifie que le point de convergence correspond bien à ce qui est annoncé par le résultat de cette section (on trouve par exemple pour (CMo) $\bar{x}_n \approx .1664$ pour l'algorithme 1 et $\bar{x}_n \approx .1684$ pour l'algorithme 2, sachant que $\bar{x}_{0,-1} \approx .1284$ et qu'on a choisi $\lambda = .04$ dans cette section). On note des oscillations importantes dans l'erreur au consensus quand on approche de la précision machine pour l'algorithme 2.

Références

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation : Numerical Methods*, Athena Scientific, 1997.

[2] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized Gossip Algorithms," *IEEE Transactions on Inform. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[3] P. Bianchi and J. Jakubowicz, "On the convergence of a multi-agent projected stochastic gradient algorithm for nonconvex optimization," *IEEE Trans. on Automatic Control*, February 2013, [online] arXiv :1107.2526v1.

[4] Patrick L. Combettes and Jean-Christophe Pesquet, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing, pp. 185–222, Springer, 2011.