Sur un algorithme de Robbins-Monro distribué

Pascal Bianchi, Gersende Fort, Walid Hachem, Jérémie Jakubowicz

Laboratoire Traitement et Communication de l'Information (LTCI)
Télécom ParisTech & CNRS
[prenom].[nom]@telecom-paristech.fr

Thème – 2 - Communication et codage (2.7 - Coopération, Diversité, Réseaux de capteurs)

Problème traité – Ce papier traite d'un algorithme de consensus de type Approximation Stochastique décentralisée pour la recherche de zéros d'une fonction. Ces travaux sont motivés par l'inférence statistique décentralisée via un réseau d'agents.
Le terme *approximation stochastique* fait référence aux algorithmes de type $\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, \xi_{n+1})$, où $\{\theta_n\}_n$ est une suite aléatoire à valeurs dans $\mathbb{R}^d$, $\{\gamma_n\}_n$ est une suite positive déterministe décroissante vers zero, et $\{H(\cdot, \xi_{n+1})\}_n$ est une suite de champs de vecteurs aléatoires, perturbations d'un champ moyen h; sous des hypothèses sur le pas γ_n et les perturbations H, cette procédure converge vers les zéros de h. Le terme *décentralisé* indique que des algorithmes d'approximation stochastiques sont exécutés localement au niveau de chaque agent. Ces agents sont disposés en réseaux et échangent de temps en temps de l'information relative à leur estimation courante. La procédure d'échanges entre agents est telle que à convergence de l'algorithme, tous les noeuds partagent la même estimée : il s'agit donc d'un algorithme de *consensus*.
L'algorithme que nous proposons consiste à laisser les agents résoudre un problème de recherche de zéros localement, via un schéma d'approximation stochastique; et échanger, à des instants aléatoires, leurs estimées suivant une étape de "gossip" [1]. A convergence, cet algorithme converge vers un consensus.

Originalité – Ce nouvel algorithme permet de résoudre des problèmes d'optimisation par une méthode distribuée, et plus généralement des problèmes de recherche de zéros d'une fonction dans le cas où cette fonction n'est pas explicite. Cet algorithme s'applique en particulier pour le calcul de l'estimateur du maximum de vraisemblance distribué.
La littérature sur l'estimation distribuée connaît un essor relativement récent. Kar *et al.* proposent dans [2] un algorithme où la dissimilarité entre agents est incorporée à la fonction cible H, ce qui leur permet de prendre en compte la quantification au prix d'une évolution plus lente vers le consensus. On trouve également dans [3] un algorithme proche de celui que nous proposons, à la différence près que dans [3], la fonction à optimiser est supposée connue de tous les agents du réseau. Ram *et al.* [4] proposent un algorithme semblable au notre et proposent une analyse de convergence sous des hypothèses assez restrictives et difficiles à vérifier en pratique.
Nous présentons aussi une analyse du comportement asymptotique de cet algorithme stochastique sous des hypothèses plus générales que [4] et facilement vérifiables. A la différence des travaux existants, nous établissons aussi un théorème central limite, ce qui nous permet d'identifier la vitesse de convergence de l'algorithme vers le consensus.

Résultats – Nous proposons un algorithme d'approximation stochastique distribué basé sur la technique de "gossip". Nous établissons sa stabilité et sa convergence vers un consensus et étudions la vitesse de convergence en démontrant un théorème central limite. Enfin, nous illustrons son fonctionnement sur un exemple d'estimation par maximum de vraisemblance distribué.

1 Introduction

Chercher les racines d'une équation de la forme $h(\theta) = 0$ est un problème récurrent que l'on rencontre par exemple dans les procédures d'optimisation de fonctions différentiables. Résoudre cette équation de façon itérative est une façon classique de procéder [5]. Lorsque h n'est pas connue de façon analytique mais que l'on ne dispose que d'une version bruitée $H(\theta,X)$, on peut utiliser l'algorithme de Robbins-Monro pour trouver les racines de h [6]. En traitement du signal, la maximisation de vraisemblance par descente de gradient en est une application typique. Dans ce cas particulier, l'objectif consiste à chercher les points stationnaires de la divergence de Kullback entre la vraie loi des observations et une famille paramétrique de lois. Dans ce cas le terme $H(\theta,X)$ correspond au gradient (par rapport à θ) de la log-vraisemblance de l'observation X sous le modèle de paramètre θ.

Dans ce papier, nous étudions des algorithmes de Robbins-Monro distribués (cf. par exemple [4]). Soit un graphe connexe : les sommets modélisent les agents et les arêtes définissent le voisinage de chaque agent i.e. l'ensemble des agents avec lesquels il peut communiquer. L'objectif est d'estimer $\theta_* \in \mathbb{R}^d$ tel que $h(\theta_*) = 0$ lorsque $h(\theta) = \sum_{i=1}^N h_i(\theta)$. Pour ce faire, chaque agent i, $i \in \{1, \cdots, N\}$, dispose d'approximations bruitées $\{H_i(\cdot,X_{n,i})\}_n$ de la fonction h_i. De plus, les noeuds sont capables de communiquer avec leurs voisins à certains instants (éventuellement aléatoires). Lorsque deux noeuds communiquent, ils s'entendent sur une valeur commune de leurs estimées qui supplante alors leurs estimées individuelles d'avant la communication. Il s'agit donc d'un algorithme qui mèle une étape d'approximation stochastique (cf. [7, 5] par exemple) et une étape de "gossip" (cf. [1]). Sous des hypothèses liées notamment à la procédure de "gossip", cet algorithme est un algorithme de consensus.

Les algorithmes de consensus ont fait l'objet de nombreux travaux dans la littérature récente de traitement du signal [8, 9]. Ils consistent à atteindre un état de consensus de façon distribuée et itérative dans le cas particulier où le consensus cherché est la valeur moyenne des quantités initiales figurant dans chaque noeud. Plus récemment, plusieurs auteurs ont formulé le problème d'approximation stochastique distribuée comme un problème de consensus [4, 8, 2].

La principale contribution de ce papier est de fournir une analyse complète de la convergence d'un algorithme de Robbins-Monro distribué : nous montrons que cet algorithme est *stable*, qu'il converge vers l'ensemble des zéros de h et nous étudions la vitesse de convergence. Ce papier généralise les travaux antérieurs de [4]. Tout d'abord, nos hypothèses sont moins restrictives; en particulier, nous ne supposons pas que la fonction objectif h est bornée. Ensuite nous énonçons des conditions simples qui garantissent la stabilité de la procédure stochastique (sans recourir à des étapes de stabilisation comme par exemple, les techniques de reprojection (voir e.g. [5])). Enfin nous étudions le comportement limite de cet algorithme sous des hypothèses sur les perturbations H_i et les champs moyens h_i facilement vérifiables.

2 Algorithme : description et analyse de convergence

2.1 L'algorithme

Soit un graphe *connexe* composé de N noeuds, $N \geq 1$. Etant donnée une suite de variables aléatoires $(X_{n,i})_{n\in\mathbb{N}}$ à valeur dans $\mathbb{R}^{m_i}$ correspondant à l'observation au temps n par le noeud i, une fonction $H_i : \mathbb{R}^d \times \mathbb{R}^{m_i} \to \mathbb{R}^d$, et une famille de matrices $N \times N$ aléatoires *doublement stochastiques* $\{W_n\}_n$, chaque noeud i construit itérativement un processus stochastique $\theta_{n,i}$ à valeurs dans $\mathbb{R}^d$. Chaque itération comprend deux étapes :

`[Étape locale]` Au temps n, chaque noeud i calcule $\tilde{\theta}_{n,i}$ par :

$$\tilde{\theta}_{n,i} = \theta_{n-1,i} + \gamma_n H_i(\theta_{n-1,i}\,; X_{n,i}) \,, \tag{1}$$

où γ_n désigne un pas déterministe positif.

`[Étape de "gossip"]` Chaque noeud i calcule une nouvelle estimée comme la moyenne pondérée des estimées voisines :

$$\theta_{n,i} = \sum_{j=1}^N w_n(i,j)\, \tilde{\theta}_{n,j} \,,$$

où $W_n := [w_n(i,j)]_{i,j=1}^N$. Lorsque $w_n(i,i) = 1$, $\theta_{n,i} = \tilde{\theta}_{n,i}$ et le noeud i n'échange avec aucun voisin. La loi des matrices W_n est telle que $w_n(i,j) = 0$ si les noeuds i et j ne sont pas connectés.

Fomulation vectorielle : Cet algorithme peut se résumer sous forme vectorielle par

$$\boldsymbol{\theta}_n = (W_n \otimes I_d)(\boldsymbol{\theta}_{n-1} + \gamma_n H(\boldsymbol{\theta}_{n-1}; X_n)) \,, \tag{2}$$

où $\otimes$ désigne le produit de Kronecker, I_d la matrice identitié $d \times d$, la fonction $H : \mathbb{R}^{dN} \times \mathbb{R}^{\sum m_i} \to \mathbb{R}^{dN}$ est donnée par

$$H(\boldsymbol{\theta}; x) := \left(H_1(\theta_1; x_1)^T, \cdots, H_N(\theta_N; x_N)^T\right)^T$$

$x = (x_1, \ldots, x_N)$ et $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$. ^{T} désigne la transposition.

2.2 Analyse de la convergence

Notations. Soit $\mathbb{1}$ le vecteur de $\mathbb{R}^N$ égal à $(1,\cdots,1)$. On note $J := (\mathbb{1}\mathbb{1}^T/N)\otimes I_d$ le projecteur orthogonal sur le sous-espace de consensus $\left\{\mathbb{1}\otimes\theta : \theta\in\mathbb{R}^d\right\}$ et $J^\perp := I_{dN} - J$ le projecteur orthogonal sur le sous-espace orthogonal à l'espace de consensus. Pour tout $\boldsymbol{\theta}\in\mathbb{R}^{dN}$, on a

$$\boldsymbol{\theta} = \mathbb{1}\otimes\langle\boldsymbol{\theta}\rangle + J^\perp\boldsymbol{\theta} \qquad \text{où} \qquad \langle\boldsymbol{\theta}\rangle := \frac{1}{N}(\mathbb{1}^T\otimes I_d)\boldsymbol{\theta}\,. \tag{3}$$

Notens que par définition de J, $\langle\boldsymbol{\theta}\rangle = (\theta_1+\cdots+\theta_N)/N$ est la moyenne des N sous-blocs de taille d du vecteur $\boldsymbol{\theta}$.

On suppose que la loi conditionnelle des v.a. $(X_{n,1},\cdots,X_{n,N},W_n)$ sachant le passé jusqu'à l'instant $(n-1)$ est donnée par

$$\mathbb{P}\left((X_{n,1},\cdots,X_{n,N})\in A, W_n\in B \mid X_{j,i},W_j, j\le n-1, i\le N\right) = \mu_{\boldsymbol{\theta}_{n-1}}(A)\,\mathbb{P}(W_1\in B)\ ,$$

où $\{\mu_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\mathbb{R}^{dN}}$ est une famille de probabilités sur $\mathbb{R}^{\sum m_i}$. Cela entraine que, étant donné le passé, les v.a. $(X_{n,1},\cdots,X_{n,N})$ et W_n sont indépendantes; et que les v.a. $\{W_n\}_n$ sont indépendantes et identiquement distribuées (i.i.d.).

Définissons le *champ moyen* $h:\mathbb{R}^d\to\mathbb{R}^d$ de l'algorithme de Robbins-Monro distribué par :

$$h(\theta) := \int \langle H(\mathbb{1}\otimes\theta;x)\rangle\ \mu_{\mathbb{1}\otimes\theta}(dx)\,. \tag{4}$$

Résultats. L'analyse du comportement asymptotique (quand $n\to+\infty$) repose sur l'existence d'une fonction de Lyapunov V pour le champ moyen h, c'est-à-dire une fonction telle que $\nabla V^T\ h\le 0$ où ∇ désigne l'opérateur de gradient. L'étude de la suite aléatoire définie par (2) nécessite tout d'abord de montrer la *stabilité* de la procédure i.e. montrer que, avec probabilité 1, la suite des estimées reste dans un compact. Ensuite, il s'agit de montrer que cette suite (stable) converge vers l'ensemble $\mathcal{L} := \{\nabla V^T\ h = 0\}$. Enfin, nous montrons que la vitesse de convergence vers $\mathcal{L}$ est liée au choix du pas γ_n. Toutes ces études nécessitent des conditions de régularité sur les fonctions H, V; sur l'ensemble $\mathcal{L}$; et des conditions sur le pas γ_n. Dans ce résumé, nous n'énonçons pas les hypothèses exactes et nous contentons d'une brève description.

Pour la procédure (2), nous établissons

la stabilité : $\limsup_n|\boldsymbol{\theta}_n| < +\infty$ avec probabilité 1. Autrement dit, pour toute suite définie par (2), il existe un compact $\mathcal{K}$ de $\mathbb{R}^{dN}$ tel que $\boldsymbol{\theta}_n\in\mathbb{R}^{dN}$.

la convergence vers $\mathcal{L}$: avec probabilité 1, la suite $\{J^\perp\boldsymbol{\theta}_n\}_n$ converge vers 0 et la suite $\{J\boldsymbol{\theta}_n\}_n$ converge vers une composante connexe de l'ensemble $\mathcal{L}$. Autrement dit, la suite $\{\boldsymbol{\theta}_n\}_n$ converge avec probabilité 1 vers l'espace de consensus $\mathcal{L}$. Ce résultat entraine aussi que lorsque $\mathcal{L}$ est un ensemble fini, la suite converge vers un point θ_* de $\mathcal{L}$.

la vitesse de convergence vers un point $\theta_*\in\mathcal{L}$: soit $\theta_*\in\mathcal{L}$ tel que $\mathbb{P}[\langle\boldsymbol{\theta}_n\rangle\to\theta_*]>0$, $\nabla h(\theta_*)$ est stable (i.e. les parties réelles de ses valeurs propres sont majorées par $-L$, $L>0$) et $Q(\theta_*) := \int\langle H(\mathbb{1}\otimes\theta_*;x)\rangle\langle H(\mathbb{1}\otimes\theta_*;x)\rangle^T\ \mu_{\mathbb{1}\otimes\theta_*}(dx)$ est définie positive. Si $\gamma_n = \gamma_0 n^{-a}$ avec $a\in(1/2,1)$ ou $\gamma_n=\gamma_0 n^{-1}$ avec $2L\gamma_0>1$, alors conditionnellement à l'évènement $[\langle\boldsymbol{\theta}_n\rangle\to\theta_*]$, on a:

$$\gamma_n^{-1/2}(\boldsymbol{\theta}_n - \mathbb{1}\otimes\theta_*) \xrightarrow{\text{loi}} \mathbb{1}\otimes Z$$

où Z est un vecteur gaussien $d\times 1$ centré de matrice de covariance Σ est l'unique solution de :

$$(\nabla h(\theta_*)^T + \zeta I_d)\Sigma + \Sigma(\nabla h(\theta^*) + \zeta I_d) = -Q(\theta^*)$$

où $\zeta = 0$ si $a\in(1/2,1)$ et $\zeta = 1/(2\gamma_0)$ si $a=1$.

Les résultats de convergence (resp. la vitesse de convergence) sont établis à l'aide des travaux de [10] (resp. [11]). Tous ces résultats requièrent la stabilité de la suite $\{\boldsymbol{\theta}_n\}_n$. Quand celle-ci n'est pas garantie par construction, elle peuit être réalisée en ajoutant à la procédure itérative de construction de $\{\boldsymbol{\theta}_n\}_n$ une étape de projection (sur des compacts "fixes" ce qui peut avoir pour effet de modifier l'ensemble des points limites; ou des compacts "croissants". Voir [5] par exemple). Dans le cas présent, nous montrons que moyennant des hypothèses convenables sur V, H et h, la procédure (2) est stable par nature.

3 Application à l'estimation par Maximum de Vraisemblance distribué

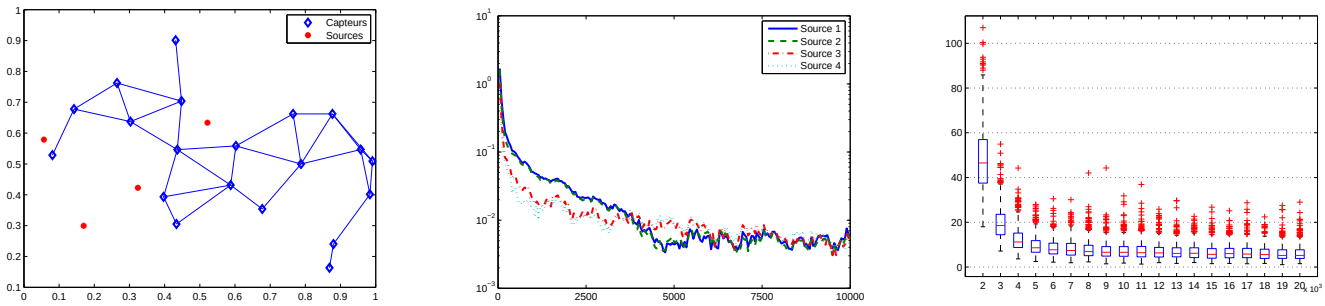
On dispose d'observations i.i.d. $\{(X_{n,1},\cdots,X_{n,N})\}_n$ de loi de densité f_* sur $\mathbb{R}^{\sum m_i}$, et d'une famille de densités $\{f(x,\theta)\}_{\theta\in\mathbb{R}^d}$ sur $\mathbb{R}^{\sum m_i}$ de la forme $f(x,\theta) = \prod_{i=1}^N f_i(x_i;\theta)$ où $f_i(\cdot;\theta)$ est une densité sur $\mathbb{R}^{m_i}$. Il s'agit d'estimer l'estimateur de maximum de vraisemblance distribué.

Dans cette application, $H_i(\theta,x) = \nabla_\theta\log f_i(x;\theta)$; $\mu_{\boldsymbol{\theta}}(dx) = f_*(x)dx$ pour tout $\boldsymbol{\theta}\in\mathbb{R}^{dN}$ et

$$h(\theta) = \frac{1}{N}\sum_{i=1}^N\int\nabla_\theta\log f_i(x;\theta)f_*(x)dx\,, \qquad \theta\in\mathbb{R}^d\,.$$

Il est facile de vérifier que la fonction $V(\theta) := \int \log f_\star(x)\ f_\star(x)dx - \int \log f(x;\theta)\ f_\star(x)dx$ est une fonction de Lyapunov. Puisque $\nabla V = -h$, on a $\mathcal{L} = \{\theta \in \mathbb{R}^d : \nabla V(\theta) = 0\}$ et les résultats énoncés en Section 2 assurent la convergence de l'algorithme de Maximum de Vraisemblance distribué vers les mêmes points limites que l'algorithme de Maximum de Vraisemblance standard (i.e. centralisé). En observant que V est la distance de Kullback entre f_* et la famille $\{f(x,\theta)\}_{\theta\in\mathbb{R}^d}$, on retrouve l'interprétation de l'estimateur du maximum de vraisemblance comme recherche de la densité dans la famille $\{f(\cdot,\theta)\}_{\theta\in\mathbb{R}^d}$ qui minimise la distance de Kullback à la vraie loi f_*.

Nous illustrons numériquement le comportement asymptotique de cet algorithme. Pour ce faire, on considère le réseau décrit Figure 1a : on dispose $N = 20$ capteurs sur le carré $[0,1] \times [0,1]$ (représentés par un cercle), qui veulent estimer la position de $D = 4$ sources (représentées par une étoile); les traits pleins représentent le système de voisinage de chaque capteur. La position des capteurs est connue (et notée z_i, $z_i \in \mathbb{C}$). On veut estimer $\theta_* = (\theta_*(1), \cdots, \theta_*(D)) \in \mathbb{C}^D$ - ou, de façon équivalente, un vecteur de $\mathbb{R}^d$ avec $d = 2D$ -, le vecteur collectant les coordonnées des D sources. Les observations sont simulées en prenant $f_* \sim \mathcal{CN}(\theta_\star, \mathrm{Diag}\left(\sigma^2_{i,1} \dots \sigma^2_{i,D}\right))$ où les variances $\sigma^2_{i,1} \dots \sigma^2_{i,D}$ sont supposées connues et données par $\sigma^2_{i,k} = 10|\theta_*(k) - z_i|^2 + 0.1$. On implémente l'algorithme (2) en prenant $\gamma_n \sim 0.1\, n^{-0.6}$. Pour l'étape de "gossip", on sélectionne aléatoirement un capteur i et un de ses voisins j, et on pose $w_n(i,j) = w_n(j,i) = 1/2$, $w_n(k,k) = 1$ pour $k \neq \{i,j\}$.
Sur la Figure 1b, on visualise la convergence vers l'espace de consensus. On définit $\Delta_n(k) := |\boldsymbol{\theta}_n(k) - \mathbb{1} \otimes \theta_*(k)|$, où $\boldsymbol{\theta}_n(k) \in \mathbb{C}^N$ est le vecteur extrait de $\boldsymbol{\theta}_n$ qui collecte les N estimées de la position de la source k. Pour chacune des sources $k \in \{1, \cdots, D\}$, on trace l'évolution de $\Delta_n(k)$ en fonction du nombre d'itérations n. Cette quantité tend vers zero, ce qui illustre la convergence vers l'espace de consensus et plus précisément, vers le maximum de vraisemblance θ_*.
Sur la Figure 1c, on visualise la vitesse de convergence donnée par le TCL : on trace $n \mapsto \gamma_n^{-1}\mathcal{E}_n$ où $\mathcal{E}_n$ est une estimation par une méthode de Monte Carlo basée sur 500 réplications, de l'erreur quadratique $\mathbb{E}\left[|\boldsymbol{\theta}_n - \mathbb{1} \otimes \theta_*|^2\right]$. On peut observer la convergence de cette quantité, ce qui confirme l'expression de la vitesse de convergence de l'erreur quadratique d'estimation.



(a) Réseau de 20 capteurs et 4 sources

(b) Pour chacune des sources k (ordonnées par abscisse croissante), tracé de $n \mapsto \Delta_n(k)$

(c) Tracé de $n \mapsto \gamma_n^{-1}\mathcal{E}_n$

Références

[1] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. on Inform. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[2] S. Kar and J.M.F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Trans. on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, 2010.

[3] Dimitri P. Bertsekas and John N. Tsitsiklis, *Parallel and distributed computation: numerical methods*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[4] S.S. Ram, A. Nedic, and V.V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Arxiv preprint arXiv:0811.2595*, 2008.

[5] H.J. Kushner and G.G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, second edition edition, 2003.

[6] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. of Mathem. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[7] A. Benveniste, M. Metivier, and Priouret P., *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, 1987.

[8] I.D. Schizas, A. Ribeiro, and G.B. Giannakis, "Consensus in ad hoc WSNs with noisy links-Part I: Distributed estimation of deterministic signals," *IEEE Trans. on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[9] S. Barbarossa and G. Scutari, "Decentralized maximum-likelihood estimation for sensor networks composed of nonlinearly coupled dynamical systems," *IEEE Trans. on Signal Processing*, vol. 55, no. 7, pp. 3456–3470, July 2007.

[10] C. Andrieu, E. Moulines, and P. Priouret, "Stability of stochastic approximation under verifiable conditions," *SIAM J. Control Optim.*, vol. 44, no. 1, pp. 283–312, 2005.

[11] M. Pelletier, "Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing," *Annals of Applied Probability*, vol. 8, no. 1, pp. 10–44, 1998.