

Approximation stochastique distribuée : Le coût de la non-bistochasticité

*Gemma MORRAL¹, Pascal BIANCHI¹, Gersende FORT¹, Jérémie JAKUBOWICZ²

¹LTCI, Telecom ParisTech & CNRS
46, rue Barrault, 75634 Paris Cedex 13, France

²SAMOVAR, Telecom Sud-Paris & CNRS
9 rue Charles Fourier, 91011 Evry, France

gemma.morral, pascal.bianchi, gersende.fort@telecom-paristech.fr
jeremie.jakubowicz@telecom-sudparis.fr

Résumé – On s’intéresse au problème de l’approximation stochastique dans un contexte distribué. L’algorithme itératif étudié se déroule en deux étapes : une étape locale d’approximation stochastique effectuée par chaque agent, et une étape de *commérag* (en anglais *gossip*) consistant en des moyennes pondérées des estimées locales. La matrice des coefficients de pondération utilisés dans l’étape de gossip est supposée stochastique par ligne. Nous relâchons l’hypothèse de bistochasticité afin d’avoir des protocoles de communication moins restrictifs et nous caractérisons la dégradation des performances qui en résulte.

1 Approximation stochastique distribuée

L’approximation stochastique (AS) est un sujet de recherche très actif depuis plus de soixante ans (voir par ex. [1, 2]). L’algorithme archétype consiste en l’itération $\theta_n = \theta_{n-1} + \gamma_n Y_n$, où $(\theta_n)_{n \geq 0}$ est une suite de paramètres vectoriels, $(Y_n)_{n \geq 0}$ est une suite de vecteurs aléatoires (v.a.) et $(\gamma_n)_{n \geq 0}$ est une suite déterministe de pas. Dans le cas des algorithmes du *gradient stochastique*, Y_n est donnée par $Y_n = -\nabla g(\theta_{n-1}) + \xi_n$ où g est une certaine fonction à maximiser et où $(\xi_n)_{n \geq 0}$ représente un bruit perturbant l’observation du gradient. Cet algorithme est centralisé en ce sens que tous les calculs sont effectués par une seule machine qui a accès à l’ensemble du vecteur aléatoire Y_n . Nous nous intéressons dans cet article au cas décentralisé. Les applications comme l’estimation statistique décentralisée dans les réseaux de capteurs, l’optimisation distribuée ou le traitement des grandes bases de données en sont un exemple.

Dans la littérature on trouve deux approches pour résoudre ce problème d’optimisation distribué. La première est l’approche *incrémentale* proposée par [3, 4, 5]. Un message contenant l’estimée actuelle est transmis de façon itérative à chaque agent du réseau qui additionne leur propre valeur basée sur son observation locale et retransmet cette mise à jour à l’agent suivant. Le problème de recherche du cycle hamiltonien dans un graphe reste un problème de classe NP. Ce travail propose une deuxième approche coopérative du calcul distribué poursuivant l’idée de [6, 7], basée sur les techniques de *consensus*. Dans ce contexte les agents mettent à jour leur estimée à partir d’une

étape locale d’approximation stochastique ; puis certains agents communiquent et modifient leur estimée locale en fonction des informations échangées. Ces méthodes de *commérag* sont connues dans la littérature anglophone comme méthodes de *gossip*. Le protocole de communication est représenté par une matrice W_n à entrées positives ou nulles avec la convention que l’élément (i, j) est nul si et seulement si les agents i et j ne communiquent pas à l’instant n . Ils supposent que les matrices W_n sont doublement stochastiques ce qui implique que pour chaque transmission de i vers j il y a en même temps une transmission de j vers i . Un exemple est étudié dans [8]. L’originalité de notre approche est d’envisager un cadre de type *broadcast* [9] ; l’agent i transmet à tous ses voisins sans attendre de retour, ce qui se traduit par la non colonne-stochasticité des W_n . Récemment, [10, 11] ont établi la convergence vers le consensus dans ce cadre de non bistochasticité.

1.1 L’algorithme distribué

Dans cet article, on considère un réseau formé de N agents, ou nœuds. Chaque agent i génère une suite d’estimées $(\theta_{n,i})_{n \geq 1}$ que nous supposons ici réelle (le cas vectoriel se traite sans plus de difficultés, sinon notations, voir [12]). A l’instant n , l’algorithme étudié s’écrit en deux étapes :

[Etape locale] L’agent i génère une estimée temporaire $\tilde{\theta}_{n,i}$ donnée par :

$$\tilde{\theta}_{n,i} = \theta_{n-1,i} + \gamma_n Y_{n,i}, \quad (1)$$

où γ_n est un pas déterministe positif et $(Y_{n,i})_{n \geq 1}$ sont les observations dont l’agent i dispose.

[Gossip] L’agent i observe les valeurs $\tilde{\theta}_{n,j}$ de certains autres

*Le travail de Gemma Morral est soutenu par la DGA (Direction Générale de l’Armement Française).

agents j et génère une moyenne pondérée :

$$\theta_{n,i} = \sum_{j=1}^N w_n(i,j) \tilde{\theta}_{n,j}, \quad (2)$$

où les $w_n(i,j)$ sont des coefficients de pondération aléatoires positifs tels que $\sum_{j=1}^N w_n(i,j) = 1$ pour tout i . Introduisons les vecteurs aléatoires $\boldsymbol{\theta}_n := (\theta_{n,1}, \dots, \theta_{n,N})^T$ et $\mathbf{Y}_n = (Y_{n,1}, \dots, Y_{n,N})^T$. En posant $W_n := [w_n(i,j)]_{i,j=1}^N$, les équations (1)-(2) deviennent :

$$\boldsymbol{\theta}_n = W_n (\boldsymbol{\theta}_{n-1} + \gamma_n \mathbf{Y}_n). \quad (3)$$

Les matrices W_n sont stochastiques par ligne, soit $W_n \mathbf{1} = \mathbf{1}$, où $\mathbf{1}$ est le vecteur-colonne dont toutes les composantes sont égales à un.

1.2 Modèles du réseau et des observations

Hypothèse 1. Il existe une famille de lois $(\mu_\theta)_{\theta \in \mathbb{R}^N}$ sur \mathbb{R}^N telles que pour tout borélien A ,

$$\mathbb{P}(\mathbf{Y}_{n+1} \in A | \mathcal{F}_n) = \mu_{\boldsymbol{\theta}_n}(A) \quad p.s.$$

où $\mathcal{F}_n = \sigma(\boldsymbol{\theta}_0, Y_1, \dots, Y_n, W_1, \dots, W_n)$. De plus, \mathbf{Y}_{n+1} et W_{n+1} sont indépendants conditionnellement à \mathcal{F}_n .

Introduisons la fonction $h : \mathbb{R} \rightarrow \mathbb{R}$:

$$h(\theta) := \frac{1}{N} \mathbf{1}^T \int \mathbf{y} \mu_{\theta \mathbf{1}}(d\mathbf{y}).$$

$h(\theta)$ représente la moyenne de $\{Y_{n,i}, i \leq N\}$ conditionnellement à l'événement "tous les agents disposent de la même estimée $\theta \mathbf{1}$ à l'instant $n-1$ ".

Soit $J := \mathbf{1}\mathbf{1}^T/N$ le projecteur orthogonal sur $\text{Vect}(\mathbf{1})$ et soit $J^\perp := I_N - J$ où I_N est la matrice identité de taille N . Pour tout $x \in \mathbb{R}^N$, posons $\bar{x} := N^{-1} \sum_{i=1}^N x_i$ et $x^\perp := J^\perp x$. On pose $W_n^\perp := J^\perp W_n$.

Hypothèse 2. On assume :

a) $(W_n)_{n \geq 1}$ est une suite i.i.d. de matrices aléatoires à entrées positives ou nulles, telles que :

$$\forall n, W_n \mathbf{1} = \mathbf{1} \quad p.s.$$

b) La matrice $\mathbb{E}(W_1)$ est colonne-stochastique :

$$\mathbb{E}(W_1)^T \mathbf{1} = \mathbf{1}.$$

c) $\rho \left(\mathbb{E} \left(W_1^{\perp, T} W_1^\perp \right) \right) < 1$,

où $\rho(M)$ désigne le rayon spectral de la matrice M .

Du point de vue de l'implémentation, l'hypothèse $W_n \mathbf{1} = \mathbf{1}$ est légère. Elle signifie que chaque agent i à l'étape (2) effectue une moyenne pondérée ($\sum_j w_n(i,j) = 1$). La plupart des travaux antérieurs supposent également que $\sum_i w_n(i,j) = 1$ pour tout j , c'est à dire que les matrices W_n sont bistochastiques. Dans cet article, nous relâchons cette hypothèse, ce qui permet d'utiliser des protocoles de gossip moins exigeants en termes de coordination du réseau. En voici un exemple concret :

Exemple 1. A chaque instant n , un agent i tiré uniformément parmi tous les agents s'éveille et transmet son estimée temporelle $\tilde{\theta}_{n,i}$ à ses voisins dans un graphe. Chaque voisin j calcule la moyenne pondérée $\theta_{n,j} = \beta \tilde{\theta}_{n,i} + (1-\beta) \tilde{\theta}_{n,j}$ où $\beta \in (0, 1)$. Les agents k qui ne sont pas dans le voisinage de i (cela inclut l'agent i lui-même) posent simplement $\theta_{n,k} = \tilde{\theta}_{n,k}$. La matrice W_n correspondant à ce protocole n'est pas bistochastique. En revanche, $\mathbf{1}^T \mathbb{E}(W_1) = \mathbf{1}^T$ (voir [9]). L'hypothèse sur le rayon spectral de $\mathbb{E}(W_1^{\perp, T} W_1^\perp)$ est satisfaite dès lors que le graphe est connecté (voir [9]).

Hypothèse 3. Les pas γ_n sont strictement positifs et satisfont $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_n^2 < \infty$ et $\lim_{n \rightarrow \infty} \gamma_{n+1}/\gamma_n = 1$.

2 Convergence presque-sûre

La convergence de l'algorithme est établie en supposant l'existence d'une fonction de Lyapounov $V : \mathbb{R} \rightarrow \mathbb{R}^+$ satisfaisant :

Hypothèse 4. Il existe $V : \mathbb{R} \rightarrow \mathbb{R}^+$ telle que :

a) V est continûment dérivable.

b) Pour tout $\theta \in \mathbb{R}$, $V'(\theta)h(\theta) \leq 0$.

c) Pour tout $M > 0$, $\{\theta \in \mathbb{R} : V(\theta) \leq M\}$ est compact.

d) L'ensemble $\mathcal{L} := \{\theta \in \mathbb{R} : V'(\theta)h(\theta) = 0\}$ est tel que $V(\mathcal{L})$ est d'intérieur vide.

Hypothèse 5. a) Il existe $C > 0$ telle que $\forall \theta \in \mathbb{R}^N$:

$$\left| \int \bar{\mathbf{y}} \mu_\theta(d\mathbf{y}) - \int \bar{\mathbf{y}} \mu_{\bar{\theta} \mathbf{1}}(d\mathbf{y}) \right| \leq C |\theta^\perp|.$$

b) $\sup_n \mathbb{E} [|\mathbf{Y}_n|^2] < \infty$.

L'hypothèse 5-a) implique que si $\boldsymbol{\theta}_n$ est proche de l'espace de consensus $\text{Vect}(\mathbf{1})$, alors l'espérance de \mathbf{Y}_{n+1} conditionnellement au passé est proche de l'espérance obtenue en remplaçant $\boldsymbol{\theta}_n$ par son projeté $\bar{\boldsymbol{\theta}}_n \mathbf{1}$. L'hypothèse 5-b) est une hypothèse de stabilité qui, combinée avec l'hypothèse 1 n'est pas aisée à vérifier a priori : ce problème de stabilité est traité plus en détail dans [11]. Dans cet article, nous nous concentrons sur les problèmes de vitesse de convergence et supposons la stabilité acquise.

Théorème 1 ([11]). Sous les hypothèses 1 à 5, $(\boldsymbol{\theta}_n)_{n \geq 0}$ converge p.s. vers l'ensemble $\{\theta \mathbf{1} : \theta \in \mathcal{L}\}$.

Les estimées $\boldsymbol{\theta}_n$ s'accordent au consensus : le vecteur des désaccords $\boldsymbol{\theta}_n^\perp$ tend vers zéro et $(\bar{\boldsymbol{\theta}}_n)_{n \geq 0}$ converge vers \mathcal{L} . Si $\mathcal{L} = \{\theta^*\}$, toutes les estimés convergent vers θ^* .

3 Fluctuations des désaccords

Nous supposons dorénavant que $\mathcal{L} = \{\theta^*\}$ pour un point $\theta^* \in \mathbb{R}$ (la généralisation est fournie dans [12]).

Hypothèse 6. a) Il existe $\delta > 0$ et $\tau > 0$ tels que :

$$\sup_{|\theta - \theta^* \mathbf{1}| \leq \delta} \int |\mathbf{y}|^{2+\tau} \mu_\theta(d\mathbf{y}) < \infty.$$

- b) Pour toute fonction f continue bornée définie sur \mathbb{R}^N , $\theta \mapsto \int f d\mu_\theta$ est continue au point θ^* .
- c) La fonction $Q : \mathbb{R}^N \rightarrow \mathbb{R}$ définie par : $Q(\theta) := \int \mathbf{y} \cdot \mathbf{y}^T \mu_\theta(d\mathbf{y})$ est continue au point θ^* .
- d) La matrice de variance-covariance du vecteur aléatoire $W_1^T \mathbf{1}$ est inversible.

Nous analysons ici les fluctuations du vecteur des désaccords renormalisé $\theta_n^\perp / \gamma_{n+1}$. En multipliant chaque membre de (3) par J^\perp , on obtient sans peine :

$$\frac{\theta_n^\perp}{\gamma_{n+1}} = \alpha_n W_n^\perp \left(\frac{\theta_{n-1}^\perp}{\gamma_n} + \mathbf{Y}_n \right) \quad (4)$$

où $\alpha_n := \gamma_n / \gamma_{n+1}$ et où l'on a utilisé le fait que $W_n^\perp \theta_n = W_n^\perp \theta_n^\perp$ (n.b. : W_n est stochastique). L'hypothèse 6-b) assure que \mathbf{Y}_{n+1} sachant \mathcal{F}_n suit approximativement la loi μ_{θ^*} lorsque n est grand : $\mathbb{E}[f(\mathbf{Y}_{n+1}) | \mathcal{F}_n] = \int f d\mu_{\theta_n}$ converge vers $\int f d\mu_{\theta^*}$ pour toute fonction f continue bornée. Rappelons aussi que par l'hypothèse 3, $\lim_n \alpha_n = 1$. Pour simplifier, on désigne par X_n la suite des vecteurs des désaccords normalisé $\theta_n^\perp / \gamma_{n+1}$. Par un raisonnement grossier mais instructif, l'équation (4) s'approche d'une chaîne de Markov dont le noyau de transition $P^*(x, f)$ est donné par $P^* = \int \mathbb{E}[f(W_1^\perp(x + \mathbf{y}))] \mu_{\theta^*}(d\mathbf{y})$, c'est à dire que P^* est le noyau de la chaîne homogène $X_{n+1} = W_{n+1}^\perp(X_n + \xi_{n+1})$ où $(\xi_n)_n$ est un processus i.i.d. de loi μ_{θ^*} , la suite $(W_n)_n$ est i.i.d. et W_{n+1} est indépendante de (X_n, ξ_{n+1}) . Des arguments rigoureux sont donnés dans [12] ; ils s'appuient sur les résultats pour des chaînes de Markov adaptées [13]. Le théorème suivant est établi dans [12].

Théorème 2. *Sous les hypothèses 1 à 6 et $\mathcal{L} = \{\theta^*\}$, P^* admet une unique probabilité invariante π^* . Pour toute fonction f continue bornée, la suite $n^{-1} \sum_{k=1}^n f(\theta_k^\perp / \gamma_{k+1})$ converge presque-sûrement vers $\pi^*(f)$ et la suite $(f(\theta_n^\perp / \gamma_{n+1}))_n$ converge en loi vers $\pi^*(f)$. Les premiers moments de la distribution π^* sont donnés par les expressions suivantes :*

$$\lim_{n \rightarrow \infty} \gamma_n^{-1} \mathbb{E}(\theta_n^\perp) = R \cdot h^* \quad (5)$$

$$\lim_{n \rightarrow \infty} \gamma_n^{-2} \text{vec} \mathbb{E}(\theta_n^\perp \theta_n^{\perp, T}) = \text{vec}(ST), \quad (6)$$

en ayant posé :

$$R := (I_N - \mathbb{E}(W_1^\perp))^{-1} \mathbb{E}(W_1^\perp)$$

$$S := (I_{N^2} - \mathbb{E}(W_1^\perp \otimes W_1^\perp))^{-1} \mathbb{E}(W_1^\perp \otimes W_1^\perp)$$

$$T := Q^{*, \perp} + R h^* \cdot h^{*, T} + h^* \cdot h^{*, T} R^T.$$

La suite $(\theta_n^\perp)_{n \geq 0}$ converge (en un certain sens) vers zéro à la vitesse γ_n , c'est à dire plus rapidement que la vitesse standard $\sqrt{\gamma_n}$ caractéristique de la plupart des algorithmes d'approximation stochastique. Par suite, les désaccords entre agents constituent une partie négligeable de l'erreur d'estimation.

4 Fluctuations de l'estimée moyenne

Hypothèse 7. *La fonction h est de classe C^2 dans un voisinage de θ^* . En outre, $H^* := -h'(\theta^*)$ est strictement positif. Enfin, on suppose que $\log(\gamma_{n-1} / \gamma_n) = o(\gamma_n)$.*

La dernière condition est satisfaite par exemple si $\gamma_n \sim \gamma_0 / n^a$ lorsque $n \rightarrow \infty$, où $a \in (0.5, 1)$. Nos résultats s'étendent au cas où $a = 1$ (voir [12]).

Théorème 3. *Sous les hypothèses 1 à 7, et $\mathcal{L} = \{\theta^*\}$, l'erreur moyenne d'estimation normalisée $\gamma_n^{-1/2} (\bar{\theta}_n - \theta^*)$ converge en loi vers $\mathcal{N}(0, \sigma_{opt}^2 + \sigma_{com}^2)$ où :*

$$\sigma_{opt}^2 = \frac{1}{2H^*} \int \bar{\mathbf{y}}^2 \mu_{\theta^*}(d\mathbf{y}),$$

$$\sigma_{com}^2 = \frac{1}{2N^2 H^*} \int \pi^*(dx) \mu_{\theta^*}(d\mathbf{y}) (x + \mathbf{y})^T C_w (x + \mathbf{y})$$

C_w désigne la matrice de covariance de $W_1^{\perp, T} \mathbf{1}$.

La preuve est omise ici (voir [12]) : elle est principalement basée sur les résultats de [13, 14]. D'après ce théorème, l'erreur moyenne d'estimation tend vers zéro à la vitesse $\sqrt{\gamma_n}$. La variance asymptotique est une somme de deux termes. Le premier σ_{opt}^2 est égal à la variance qui serait obtenue dans un contexte centralisé classique. Le second σ_{com}^2 représente l'excès de variance inhérent au contexte distribué. On note que $W_n^{\perp, T} \mathbf{1} = \mathbf{1}$ dans le cas où la matrice W_n est stochastique par colonnes. Dans ce cas, C_w est nulle et $\sigma_{com}^2 = 0$. Ainsi, l'algorithme possède les mêmes performances qu'un algorithme centralisé dans le cas où les matrices W_n utilisées sont bistochastiques (voir [11]). En revanche, des protocoles tels que celui de l'exemple 1 ne satisfont pas cette hypothèse. Dans un tel cas, σ_{com}^2 quantifie l'excès de variance.

4.1 Exemple Numérique

Dans cet exemple, $\theta_{n,i}$ est à valeur dans \mathbb{R}^2 : il collecte les coordonnées d'un point. L'objectif est d'estimer la position d'une cible $\theta^* = (a, b)$; $a = b = 1$. Le réseau est composé de $N = 10$ agents ; à l'instant n , chaque agent i dispose de l'observation $\mathbf{Y}_{n,i} = -(\theta_{n,i} - \theta^*) + \xi_{n,i}$ où les v.a. $(\xi_{n,i}, n \geq 0, i \leq N)$ sont i.i.d. de loi $\mathcal{N}_2(0, \sigma^2 Id_2)$, $\sigma^2 = 2$. Par suite, μ_{θ_n} est la distribution $\mathcal{N}_{2N}(\mathbf{1} \otimes \theta^* - \theta_n, 2Id)$. On prend $\gamma_n = 1/n^{0.7}$. La fonction $h : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ est donnée par $h(\theta) = (\theta^* - \theta)$. L'hypothèse 4 est vérifiée avec $V(\theta) = \|\theta^* - \theta\|^2$ et $\mathcal{L} = \{\theta^*\}$. Deux protocoles de comméragement sont envisagés : le protocole *pairwise* (voir [8]) pour lequel les matrices W_n sont bistochastiques ; et le protocole *broadcast* décrit dans l'exemple 1 avec $\beta = 1/2$ qui n'est pas bistochastique. Pour comparaison, on considère aussi un algorithme d'optimisation centralisé pour la résolution de $\arg\min_{\theta} \|\theta - \theta^*\|^2$. On illustre tout d'abord la convergence des algorithmes : sur la Figure 1, on représente la position de la cible (\blacklozenge) et la valeur moyenne des estimées $1/N \sum_{i=1}^N \theta_{n,i}$ à différentes itérations $n = 1, 2, 50, 100, 1000, 10\,000$ et $60\,000$ (\times). On remarque que l'erreur d'estimation $\|\theta_n - \theta^*\|$ reste inférieure à 0.01 après les 3 300 itérations pour le cas centralisé et que l'erreur d'estimation moyenne $1/N \sum_{i=1}^N \|\theta_{n,i} - \theta^*\|$ reste inférieure à 0.01 après les 3 600 itérations et les 13 000 itérations pour les cas pairwise et broadcast respectivement.

La Figure 2 illustre le Théorème 2. Dans cet exemple, $h^* = 0$, $T = Q^{*, \perp} = \sigma^2 J^\perp$; dans le cas pairwise, $R = S =$

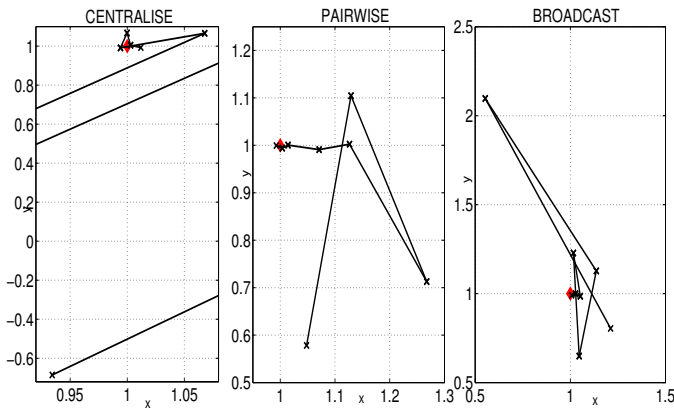


FIGURE 1: Position de la cible θ^* (\blacklozenge); valeur moyenne (sur les agents) de l'estimée à différentes itérations (\times).

$(N - 2)J^\perp$ tandis que dans le cas broadcast, $R = J^\perp$ et $S = 1/3J^\perp$. On considère 500 réalisations indépendantes de l'algorithme, itéré sur 60 000 itérations : la Figure 2 représente les boxplots des 500 valeurs de l'abscisse du premier agent $\theta_{n,1}^\perp/\gamma_{n+1}$ lorsque $n = 60\,000$. On constate que les résultats sont bien meilleurs pour l'algorithme broadcast : le calcul théorique des moments d'ordre 1 et 2 de la loi limite π^* donne $(0, 0.6)$ pour le broadcast et $(0, 14.4)$ pour le pairwise ; empiriquement, nous obtenons $(-0.01, 0.66)$ pour le broadcast et $(-0.4, 14.39)$ pour le pairwise.

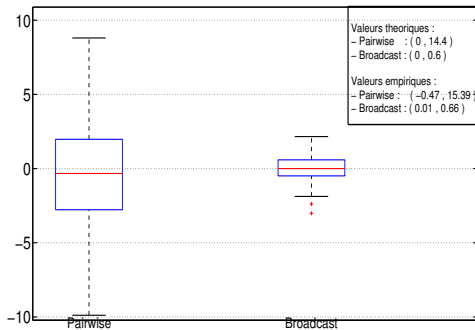


FIGURE 2: Boxplot de l'abscisse de la suite des désaccords du premier agent $\theta_{n,1}^\perp/\gamma_{n+1}$ décrite par (4) lorsque $n = 60\,000$.

La Figure 3 illustre le Théorème 3 sur le comportement asymptotique de la suite vectorielle $\Delta_n = \gamma_n^{-1/2}(\bar{\theta}_n - \theta^*)$: on trace l'histogramme des 500 valeurs de l'abscisse de Δ_n lorsque $n = 60\,000$. Dans cet exemple, $\sigma_{\text{opt}}^2 = \sigma^2/(2N)$ et $\sigma_{\text{com}}^2 = 1/(2N)(\text{vec}(C_w)^T \text{vec}(ST) + \text{tr}(C_w Q^{*,\perp}))$ avec la matrice de covariance $C_w = \beta^2 N J^\perp$; ces valeurs théoriques sont reportées sur la Figure 3. Les variances du cas centralisé et du cas pairwise correspondent à σ_{opt}^2 ; empiriquement, on trouve qu'elles valent respectivement 0.115 et 0.078 (valeur théorique 0.1). La variance du cas broadcast est la somme des deux termes de variance $\sigma_{\text{opt}}^2 + \sigma_{\text{com}}^2$; empiriquement, on trouve 0.42 (valeur théorique 0.39). Ce graphe montre que la performance est la même pour l'algorithme centralisé et l'algorithme distribué

avec des matrices W_n bistochastiques ; tandis que l'on peut observer un terme supplémentaire dans la variance limite lorsque cette condition de bistochasticité n'est plus vérifiée.

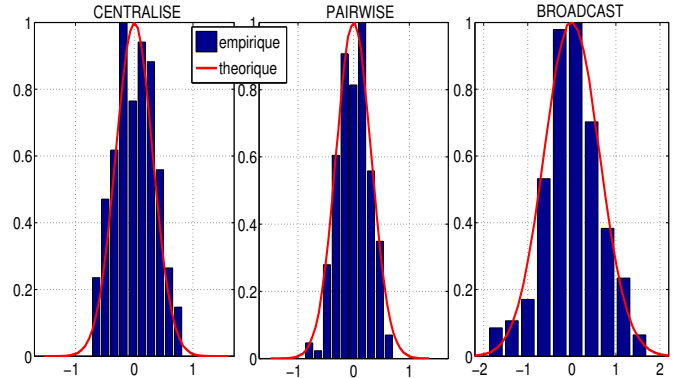


FIGURE 3: Histogramme de l'erreur moyenne d'estimation de l'abscisse de Δ_n . On superpose la loi Normale qui résulte du Théorème 3.

Références

- [1] A. Benveniste, M. Metivier, and Priouret P. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, 1987.
- [2] H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.
- [3] M. Rabbat and R. Nowak. Distributed Optimization in Sensor Networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27. ACM, 2004.
- [4] A. Nedic and D.P. Bertsekas. Incremental Subgradient Methods for Non-differentiable Optimization. *SIAM Journal of Optimization*, 12(1) :109–138, 2001.
- [5] S.S. Ram, A. Nedic, and V.V. Veeravalli. Incremental Stochastic Subgradient Algorithms for Convex Optimization. *SIAM Journal on Optimization*, 20(2) :691–717, 2009.
- [6] J. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, Massachusetts Institute of Technology, 1984.
- [7] A. Nedic, A. Ozdaglar, and P.A. Parrilo. Constrained Consensus and Optimization in Multi-Agent Networks. *IEEE Trans. on Automatic Control*, 55(4) :922–938, April 2010.
- [8] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized Gossip Algorithms. *IEEE Trans. on Inform. Theory*, 52(6) :2508–2530, 2006.
- [9] T.C. Aysal, M.E. Yildiz, A.D. Sarwate, and A. Scaglione. Broadcast Gossip Algorithms for Consensus. *IEEE Trans. on Signal Processing*, 57(7) :2748–2761, 2009.
- [10] A. Nedic. Asynchronous Broadcast-Based Convex Optimization Over a Network. *IEEE Trans. on Automatic Control*, 56(6) :1337–1351, June 2011.
- [11] P. Bianchi, G. Fort, and W. Hachem. Performance of a Distributed Stochastic Approximation Algorithm. *IEEE Trans. on Inform. Theory (submitted) arXiv1203.1505 [math.OA]*, 2012.
- [12] P. Bianchi, G. Fort, and G. Morral. Distributed Stochastic Approximation : The Price of Non-double Stochasticity. In progress, 2013.
- [13] G. Fort, E. Moulines, and P. Priouret. Convergence of Adaptive and Interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6) :3262–3289, 2012.
- [14] G. Fort. Central limit theorems for stochastic approximation algorithms. Technical report, 2012.