# Convergence des méthodes de gradient stochastique sur les variétés riemanniennes

Silvère Bonnabel

Centre de Robotique, Mines ParisTech, Paris, France

silvere.bonnabel@mines-paristech.fr

**Résumé –** On propose un algorithme de gradient stochastique en ligne sur les variétés riemanniennes. Le cadre géométrique envisagé permet d'unifier plusieurs algorithmes célèbres. Par ailleurs la méthode proposée peut être appliquée à de nouveaux problèmes d'apprentissage en ligne, et de traitement de certaines images spécifiques dont les pixels sont des matrices. De plus, on fournit une preuve de convergence vers un point critique de la fonction de coût moyennée sous des hypothèses faibles. On étend les résultats de convergence connus dans le cas euclidien au cas riemannien. La principale différence est que la longueur du pas doit varier en fonction de la courbure et des non-linéarités du gradient.

**Abstract –** In this paper an online stochastic gradient descent algorithm on Riemannian manifolds is proposed. The geometric framework allows to unify several known algorithms, and opens the way to new applications in online learning. Moreover, the known results of convergence to a critical point of the averaged cost function under weak assumptions are extended to the Riemannian case. The main difference is the use of an adaptative stepsize which takes into account some specific non-linearities of the problem.

## 1 Introduction

Les méthodes d'approximations stochastiques sont des méthodes d'optimisation basées sur des approximations de la fonction de coût. Elles sont particulièrement adaptées au problèmes d'apprentissage ou identification en ligne, et suivi d'un signal qui évolue lentement (tracking). Elles ont également connu un regain d'intérêt récent dans des problèmes batch avec l'explosion des problèmes de large dimension, dans lesquels les échantillons à traiter ne peuvent pas être gérés numériquement dans leur globalité. La méthode la plus simple est le gradient stochastique qui a été intensivement utilisé pour sa simplicité en filtrage adaptatif (LMS), traitement du signal [2] et en apprentissage où la méthode a été popularisée notamment par L. Bottou [5].

Dans cet article nous proposons un algorithme de descente de gradient stochastique sur les variétés riemanniennes et nous fournissons une preuve de convergence. La méthode est applicable à divers problèmes d'optimisation et de régression (en ligne) dans lesquels certaines contraintes spécifiques peuvent être vues comme l'appartenance du paramètre à une variété riemannienne, et auxquels s'appliquent la machinerie de l'optimisation sur les variétés (voir e.g. [1, 6]).

Les applications potentielles en traitement du signal sont nombreuses. En effet on peut penser aux problèmes d'identification et de suivi en ligne lorsque le paramètre peut être vu comme élément d'une variété. Par exemple l'algorithme fondamental d'Oja [8] pour l'identification et le suivi de sous-espace dominant peut être vu comme un cas particulier du présent article, puisqu'il s'interprète comme gradient stochastique sur la variété de Stiefel. On pense aussi aux problèmes de régression avec des contraintes de rang faible, voir [7] où la convergence du gradient est seulement conjecturée et pour lequel le présent article fournit une preuve de convergence. Finalement on pense aux applications pour des images dont les pixels sont des points d'une variété, notamment en imagerie médicale (voir e.g. [9]) et en traitement du signal radar [4] où le présent algorithme pourrait permettre de segmenter les images par clustering (K-means). On peut aussi penser au calcul de moyennes sur des variétés. Par exemple, l'article récent [3] utilise un algorithme de gradient stochastique pour trouver le barycentre d'une mesure sur une variété et montre la convergence (pour la fonction de coût correspondante).

Il est à noter enfin que ces travaux peuvent être évidemment reliés aux méthodes de la géométrie de l'information. En effet, lorsque le paramètre à identifier peut être vu comme paramètre d'un modèle statistique, la métrique de Fisher fournit une métrique riemannienne. Notre algorithme coincide alors avec le très célèbre algorithme de gradient naturel introduit par Amari [2], et dont l'auteur a montré les propriétés statistiques, à savoir qu'il converge en probabilité, et qu'il est efficace pour une certaine longueur de pas (i.e. la matrice de covariance de l'erreur tend vers la borne de Cramer-Rao). Cet algorithme a déjà été utlisé avec succès pour le problème de la séparation de sources.

## 2 Descente de gradient stochastique sur les variétés riemanniennes

En s'inspirant de la formulation du problème par [5] (dont la lecture est recommandée), on considère le problème d'optimi-

sation qui consiste à *minimiser* la fonction de coût :

$$C(w) = \mathbb{E}_z Q(z,w) = \int Q(z,w) dP(z)$$

où $w \in \mathcal{M}$ est un paramètre de minimisation appartenant à une variété riemannienne $\mathcal{M}$, et $z$ est un évènement appartenant à un espace mesurable $\mathcal{Z}$, modélisé comme une variable aléatoire distibuée selon une loi inconnue $dP$. $Q(z,w)$ est la fonction d'erreur (loss fonction), qui peut être vue comme une approximation de la fonction de coût évaluée sous l'évènement $z$.

## 2.1 Descente de gradient stochastique

On suppose la fonction de coût dérivable trois fois. L'algorithme classique de gradient stochastique dans le cas euclidien $\mathcal{M} = \mathbb{R}^n$ s'obtient en tirant un évènement $z_t$ selon la loi de probabilité associée et en écrivant $w_{t+1} = w_t - \gamma_t H(z_t, w_t)$ où $\mathbb{E}_z H(z,w) = \nabla C(w)$. Comme $C$ n'est pas convexe dans de nombreuses applications, on ne peut pas vraiment espérer mieux dans le cas général qu'une convergence presque sure de $C(w_t)$, et de $\nabla C(w_t)$ vers zéro. C'est en effet ce que l'on peut montrer sous une série d'hypothèses. Dans le cas d'une variété riemannienne on propose de remplacer la mise à jour habituelle par la formule intrinsèque

$$w_{t+1} = \exp_{w_t}(-\frac{\gamma_t}{f(w_t)} H(z_t, w_t)) \tag{1}$$

où $\exp_w$ est l'application exponentielle au point $w$, c'est à dire que si $v$ est un vecteur tangent en $w$, $\exp_w(v)$ est le point de la variété qui se situe sur la géodésique d'origine $w$, dans la direction $v$, à une distance $\|v\|$ du point $w$; et où le gradient de la fonction de coût évaluée en $z_t$ est défini comme le gradient au sens de la métrique choisie sur la variété, c'est donc un vecteur tangent en $w$. Soit $D(w_1, w_2) = d^2(w_1, w_2)$ la distance riemannienne au carré. On montre la convergence sous les hypothèse suivantes, qui peuvent être vues comme une généralisation de celles de [5].

1. $\sum \gamma_t^2 < \infty$ et $\sum \gamma_t = +\infty$.
2. Il existe $w^* \in \mathcal{M}$ et $S > 0$ tel que
$$\inf_{D(w,w^*)>S} \langle \exp_w^{-1}(w^*), \nabla C(w) \rangle > 0$$
i.e. le gradient pointe vers $w^*$ quand $w$ s'éloigne trop de ce point.
3. Il existe $E > S$ tel que le gradient est borné dans une région autour de $w^*$, i.e.
$$\forall z \ \|H(z,w)\| \leq A \text{ pour } D(w,w^*) \leq E$$
4. La courbure sectionnelle est partout minorée par $\kappa < 0$ et le rayon d'injectivité est non-nul.
5. $\mathbb{E}_z(\|H(z,w)\|^2)$ et $\mathbb{E}_z(\|H(z,w)\|^3)$ sont finis de sorte qu'on peut définir la fonction $f : \mathcal{M} \mapsto \mathbb{R}$ dans (1) comme une fonction continue bornée inférieurement par 1, telle que
$$f(w)^2 \geq \max\{1, \mathbb{E}_z\big(\|H(z,w)\|^2(1 + \sqrt{|\kappa| D(w,w^*)} + \|H(z,w)\|\sqrt{|\kappa|})\big)\}$$

## 2.2 Schéma de la preuve

On montre que $C(w_t)$ converge presque sûrement et $\nabla C(w_t)$ converge presque sûrement vers 0. La principale difficulté de l'adaptation de la preuve au cas riemannien est de montrer que les trajectoires rejoignent presque sûrement un compact fixé à l'avance. En effet le théorème de Pythagore n'est plus vrai sur une variété et l'on peut s'éloigner beaucoup d'un point en suivant une géodésique pourtant proche de la géodésique qui nous relie à ce point. Les effets de courbure peuvent alors annihiler la décroissance polynomiale du pas et conduire à des trajectoires non bornées. L'originalité de la démonstration consiste donc à remplacer l'hypothèse habituelle de croissance au plus linéaire du gradient par l'hypothèse très générale 5 qui tient compte des non-linéarités de la formulation riemannienne ainsi que des effets de courbure, et à modifier le pas $\gamma_t$ en fonction de $w_t$.

## 2.3 Compacité des trajectoires

Nous montrons dans un premier temps que toutes les trajectoires restent bornées dans un compact. C'est la partie la plus dure dans le cas non-euclidien pour les raisons évoquées ci-dessus. Considérons l'update (1). Soit

$$h_t = \max(E, D(w_t, w^*))$$

Nous allons montrer que cette fonction converge p.s. vers $E$. On s'appuie sur l'expansion de Taylor qui nous fournit l'inégalité:

$$\begin{aligned} h_{t+1} - h_t \leq & -2\frac{\gamma_t}{f(w_t)} \langle H(z_t, w_t), \exp_{w_t}^{-1}(w^*) \rangle \\ & + (\frac{\gamma_t}{f(w_t)})^2 \|H(z_t, w_t)\|^2 k_1 \end{aligned} \tag{2}$$

où $k_1$ est une borne supérieure sur le hessien de $D(\cdot, w^*)$ (au sens riemannien) le long de la géodésique reliant $w_t$ à $w_{t+1}$. Mais si $\mathcal{M}$ est une variété bornée inférieurement par $\kappa < 0$ un résultat sans doute connu et écrit par Cordero-Erausquin, McCann et Schmuckenschläger en 2001 dit que

$$\begin{aligned} \nabla_w^2 (D(w,w^*)/2) &\leq \frac{\sqrt{|\kappa| D(w,w^*)}}{\tanh(\sqrt{|\kappa| D(w,w^*)})} \\ &\leq \sqrt{|\kappa| D(w,w^*)} + 1 \end{aligned}$$

Or par inégalité triangulaire $\sqrt{D(w_{t+1}, w^*)} \leq \sqrt{D(w_t, w^*)} + \|H(z_t, w_t)\|$ si l'on suppose $t$ assez grand pour que $\gamma_t \leq 1$ et puisque $f(w_t) \geq 1$. Cela nous fournit un majorant $k_1$ ne dépendant que de $w_t$. Si $F_t$ est la suite croissante des $\sigma$-algèbres engendrée par les variables réalisées (juste avant) le temps $t$:

$$F_t = \{z_0, \cdots, z_{t-1}, w_0, \cdots, w_t, \gamma_0, \cdots, \gamma_t\}$$

on a

$$\mathbb{E}[(\frac{\gamma_t}{f(w_t)})^2 \|H(z_t, w_t)\|^2 k_1 | F_t] = (\frac{\gamma_t}{f(w_t)})^2 \mathbb{E}_z\big(k_1 \|H(z_t, w_t)\|^2\big)$$

En conditionnant (2) à $F_t$, puis en utilisant l'hypothèse 5 on a

$$\mathbb{E}[h_{t+1} - h_t | F_t] \leq -2\frac{\gamma_t}{f(w_t)} \langle \nabla C(w_t), \exp_{w_t}^{-1}(w^*) \rangle + \gamma_t^2$$

Comme dans la preuve habituelle [5] on peut enlever le premier terme puisque

- Soit $D(w_t, w^*)$ et $D(w_{t+1}, w^*)$ sont plus petits que $E$.
- Soit $D(w_t, w^*) > E$ et l'hypothèse 2 assure que le terme est négatif.
- Si $D(w_t, w^*) < E$ et $D(w_{t+1}, w^*) > E$ pour $t$ suffisamment grand $\gamma_t$ est petit et l'hypothèse 3 assure que $D(w_t, w^*) > S$.

Et donc *in fine* $\mathbb{E}(h_{t+1} - h_t|F_t) \leq \gamma_t^2$ ce qui nous ramène à la preuve de [5]. En effet en sommant la dernière inégalité on voit que la somme des variations positives du processus non-négatif $h_t$ est bornée, et cela implique par un théorème de Fisk (1965) que c'est une quasi-martingale qui converge presque sûrement vers une valeur. Cette valeur ne peut être que E. En effet notons $T = 2\sum_{t_0}^{\infty} \frac{\gamma_t}{f(w_t)} \langle \nabla C(w_t), \exp_{w_t}^{-1}(w^*) \rangle$. L'inégalité ci-dessus se ré-écrit

$$T \leq \sum_{t_0}^{\infty} \gamma_t^2 - \sum_{t_0}^{\infty} \mathbb{E}[h_{t+1} - h_t|F_t]$$

et puisque $h_t$ est une quasi-martingale (par le même théorème), on a presque sûrement (par application de la définition)

$$T \leq |\sum_{t_0}^{\infty} \gamma_t^2| + \sum_{t_0}^{\infty} |\mathbb{E}[h_{t+1} - h_t|F_t]| < \infty$$

Prenons maintenant une trajectoire échantillon pour laquelle $h_t$ converge vers $E' > E$. Ce fait est incompatible avec la borne ci-dessus, car d'après l'hypothèse 2 pour $t_0$ suffisamment grand on a $\langle \nabla C(w_t), \exp_{w_t}^{-1}(w^*) \rangle > \epsilon$ avec $\epsilon > 0$. Cela est incompatible avec l'hypothèse 1 qui stipule que la somme des $\gamma_t$ est infinie. (Le terme $f(w_t)$ n'influe pas car $f$ est continue et puisque $h_t$ converge, $w_t$ reste dans un compact).

### 2.4 Convergence de l'algorithme

Une fois ce point montré, toutes les fonctions continues du paramètre peuvent être bornées comme dans la preuve [5], et la modification sur le pas devient transparente puisqu'alors $\gamma_t/k_3 \leq \gamma_t/f(w_t) \leq \gamma_t$. La preuve se conclut sans difficulté avec les mêmes arguments que dans le cas euclidien. En effet, posons

$$g_t = C(w_t) \geq 0$$

On a l'inégalité issue du développement de Taylor

$$\begin{aligned} g_{t+1} - g_t \leq & -2\frac{\gamma_t}{f(w_t)} \langle H(z_t, w_t), \nabla C(w_t) \rangle \\ & + (\frac{\gamma_t}{f(w_t)})^2 \|H(z_t, w_t)\|^2 k_2 \end{aligned} \tag{3}$$

où $k_2$ est une borne sur le hessien Riemannien de $C$ dans le compact. En passant à l'espérance conditionnelle

$$\mathbb{E}(g_{t+1} - g_t|F_t) \leq -2\frac{\gamma_t}{f(w_t)} \|\nabla C(w_t)\|^2 + {\gamma_t}^2 k_2$$

en vertu de l'hypothèse 5 sur $f(w)$. Une fois de plus cela montre que $g_t = C(w_t)$ est une quasi martingale en appliquant les étapes et le théorème vus plus haut. Pour montrer que le gradient tend vers zéro presque sûrement, on utilise le fait que $g_t$ est une quasi-martingale dans l'inégalité ci-dessus, et on obtient (par le même raisonnement exactement que celui de la sous-section précédente) que presque sûrement

$$\sum_{1}^{\infty} \gamma_t \|\nabla C(w_t)\|^2 < \infty$$

puisque la fonction $f$ est bornée sur le compact. On définit ensuite la fonction

$$p_t = \|\nabla C(w_t)\|^2$$

et l'on a l'inégalité basée sur le développement de Taylor suivante en bornant les diverses différentielles de $C$ dans le compact

$$\mathbb{E}(p_{t+1} - p_t|F_t) \leq 2\gamma_t \|\nabla C(w_t)\|^2 k_3 + {\gamma_t}^2 k_2 k_4$$

ce qui implique, avec toujours les mêmes arguments, la convergence presque sûre de $p_t$ vers une valeur qui ne peut être que nulle en raison du résultat plus haut relatif à la somme des $p_t$.

## Références


[1] P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2007.

[2] S.I. Amari. *Natural gradient works efficiently in learning*. Neural Computation, MIT Press, 1998.

[3] M. Arnaudon, C. Dombry, A. Phan, and Le Yang. Stochastic algorithms for computing means of probability measures. *Preprint*, 2010.

[4] F. Barbaresco. Innovative tools for radar signal processing based on cartan's geometry of symmetric positive-definite matrices and information geometry. In *IEEE Radar Conference*, 2008.

[5] L. Bottou. *Online Algorithms and Stochastic Approximations*. Online Learning and Neural Networks, Edited by David Saad, Cambridge University Press, 1998.

[6] T.A. Arias Edelman, A. and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[7] G. Meyer, S. Bonnabel, and R. Sepulchre. Regression on fixed-rank positive semidefinite matrices: a riemannian approach. *In Press. Journal of Machine Learning Reasearch (JMLR). http://arxiv.org/abs/1006.1288*, 2011.

[8] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927 – 935, 1992.

[9] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66:41–66, 2006.