Master 2 Internship proposal :
Asymptotic behavior of stochastic approximation algorithms

Tutor : Pascal Bianchi, Prof., Telecom Paris, Palaiseau, France, email : prenom.nom@telecom-paris.fr

Description : In order to illustrate what are stochastic approximation algorithms, consider the well-known stochastic gradient descent (SGD). The SGD is used to find local minimizers of a function $F : \mathbb{R}^d \to \mathbb{R}$. It generates iterates of the form :

$$x_{n+1} = x_n - \gamma_n \nabla F(x_n) + \gamma_n \xi_n \,,$$

where γ_n is a positive step size, and ξ_n is a random perturbation, with zero expectation, which is due to the fact that the gradient can only be computed up to some unknown random approximation error. The map ∇F is the gradient of F. Motivated by applications to neural nets, there has been an increasing interest towards the case where F is no longer differentiable. In that case, the gradient ∇F is no longer defined, and must be replaced by an element of the so-called Clarke subdifferential of F, denoted by ∂F :

$$x_{n+1} \in x_n - \gamma_n \partial F(x_n) + \gamma_n \xi_n \,,$$

where $\partial F : \mathbb{R}^n \to 2^{\mathbb{R}}$ is multivalued function, that is, $\partial F(x)$ is a subset of $\mathbb{R}$ for every x. This algorithm is a special case of the so-called stochastic approximation algorithms, which are ubiquitous in optimization, game theory and reinforcement learning. Generically, they write :

$$x_{n+1} \in x_n + \gamma_n H(x_n) + \gamma_n \xi_n \,,$$

where $H : \mathbb{R}^n \to 2^{\mathbb{R}}$ is some multivalued function. Assuming that $\gamma_n \to 0$ as $n \to 0$, the almost sure convergence of the sequence (x_n) can be studied using the so-called Differential Inclusion method [1]. The idea is to establish that the iterates (x_n) shadow the behavior of a solution $x(t)$ to the differential inclusion :

$$\frac{dx(t)}{dt} \in H(x(t)) \,. \tag{1}$$

More precisely, [1] shows that the set of accumulation points of the sequence (x_n), denoted by $\mathrm{acc}(x_n)$, is included in a so-called Internally Chain Transitive (ICT) set of the differential inclusion (1). The definition of an ICT set is involved, and is not provided here. In the simple SGD case ($H = -\partial F$), ICT sets can be characterized : they are contained in the set of critical points of F [2]. However, beyond this simple case, strange ICT sets are likely to pop up.

Very recently, [4] introduced the useful concept of *essential accumulation points*. A point $x \in \mathbb{R}^d$ is an essential accumulation point of the sequence (x_n) if, for every neighborhood U of x,

$$\limsup_{n\to\infty} \frac{\sum_{i \le n : x_i \in U} \gamma_i}{\sum_{i \le n} \gamma_i} > 0 \,.$$

The idea is the following : instead of looking at every accumulation point, one only looks at the accumulation points x for which the sequence (x_n) spends a non-negligible proportion of time in the neighborhood of x. In a recent work, we proved that the essential accumulation points are included in the so-called Birkhoff Center (BC) of the differential inclusion (1), which is the set of all recurrent points (recall that a point x is said recurrent, if there exists a solution to (1) which returns infinitely often in any neighborhood of x). This findings are interesting in the case where the BC is easy to characterize, whereas ICT sets are not. In such cases, the new concept of essential accumulation point would be revealed useful.

The goal of the internship is to review the literature in several applications (in particular Game theory, Generative Adverserial Networks, Reinforcement Learning), in order to identify the situations where the ICT sets are difficult to characterize, and to analyze the structure of the essential accumulation set. To this end, an exhaustive state of the art of applications of stochastic approximation methods is necessary.
Several research projects are likely to be proposed in the continuity of this research topic. One of them is to characterize the convergence rate of stochastic approximation methods.

Candidacy :The candidate should have a strong background in probability theory. Please send your resume, along with your grades, to the email address provided at the top of this document.

Références

[1] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control Optim.*, 44(1) :328–348 (electronic), 2005.

[2] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Found Comput Math*, (20) :119–154, 2020.

[3] M. Faure and G. Roth. Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems. *Stoch. Dyn.*, 13(1) :1250011, 23, 2013.

[4] , J. Bolte and E. Pauwels and R. Rios-Zertuche. Long term dynamics of the subgradient method for Lipschitz path differentiable functions. arXiv preprint arXiv :2006.00098.