Sujet de Thèse : Deux problèmes d'Apprentissage Profond sous l'angle de la Théorie des Jeux

February 11, 2019

1 Encadrant

Encadrants de la thèse :

- Yann Chevaleyre, Professeur, LAMSADE - UMR 7243, yann.chevaleyre@dauphine.fr

Ecole doctorale : Ecole doctorale de Dauphine - Université Paris Dauphine
Laboratoire d'accueil : LAMSADE - Université Paris Dauphine

2 Contexte

De nombreux problèmes d'apprentissage profond peuvent être formulés comme des jeux à deux joueurs.

Parmi ces nombreux problèmes, le plus important est sans doute l'apprentissage des *Réseaux Adversariaux Génératifs* (GANs [1]). Les GANs sont usuellement modélisés comme des problèmes min-max. Un GAN est constitué de deux réseaux de neurones : un générateur et un discriminateur. Lorsqu'on les utilise dans le cadre de l'apprentissage non-supervisé, on fourni aux réseaux des données réelles, et après entrainement, le générateur devra être capable de produire de nouvelles données synthétiques extrêmement proches (en distribution) des données réelles fournies. Plus précisément, le générateur est entrainé à produire des données synthétiques dont la distribution est indistinguable de la distribution des données fournies au GAN, aux yeux du discriminateur. A contrario, le discriminateur est entraîné à distinguer le mieux possible les données synthétiques des données réelles. En ce sens, c'est un jeu à somme nulle. Les GANs ont connu un immense succès dans la communauté d'apprentissage, et de très nombreux algorithmes leur sont dédiés.

Un autre problème majeur d'apprentissage qui peut être formulé comme un jeu à deux joueur est le problème des *attaques adversariales* et de la *conception de réseaux robustes à ces attaques* [4]. Il a été découvert récemment qu'en appliquant une petite perturbation à un réseau de neurone, on peut aisément l'induire en erreur sur sa tâche de prédiction. Ce problème peut être vu comme un jeu à deux joueurs ou l'attaquant cherche à concevoir une perturbation spécifique pour tromper le classifieur, alors que ce dernier cherche à résister à l'attaquant. Résister aux attaques adversarial est un enjeu immense, en particulier pour toutes les applications sensibles (par exemple les voitures autonomes).

3 Résumé du projet de thèse

La thèse pour laquelle nous demandons un financement au sein de l'école doctorale de Dauphine aura pour but d'analyser ces problèmes d'apprentissage sous l'angle de la *théorie des jeux,* et de concevoir des algorithmes plus efficaces pour résoudre ces problèmes.

Jusqu'à présent, la grande majorité des travaux de recherche appliqués aux GANs et aux attaques adversariales s'est concentré sur la recherche d'un équilibre pur (non mixte). Certains travaux très récents ont abordé succinctement l'étude des équilibres mixtes, mais les solutions apportées sont très loin d'être implémentables en pratique [3]. Dans

cette thèse, étudierons principalement les stratégies mixtes, et les équilibres de Nash mixtes (ou autres) associés, ainsi que les algorithmes permettant d'atteindre ces équilibres (de façon approchée). Plus précisément, les objectifs de la thèse seront de:

1. **Comprendre la nature des jeux associés aux GANs et aux attaques adversariales. Etablir les conditions générales sous lesquelles le jeu admet un équilibre de Nash pur.**

   Il s'agira d'étudier la topologie et la géométrie du problème pour déterminer si la nature du jeu est connue (jeux de potentiel, jeux stables etc...) ou s'il n'est pour l'instant pas défini. Dans le cas où il est connu, nous pourrons nous appuyer sur les propriétés déjà démontrées en théorie des jeux pour déterminer l'existence d'un équilibre de Nash unique, et dans le cas contraires l'enjeu sera de déceler les spécificités du nouveau jeu pour ensuite tenter de démontrer l'existence de cet équilibre.

2. **Etudier les stratégies mixtes associés à ces jeux, et formaliser les équilibres de Nash mixtes (et autres). Etudier le "prix de la pureté".**

   De même qu'il existe un *prix de l'anarchie*, nous élaborerons un *prix de la pureté*, comme étant la différence de qualité entre une solution mixte et une solution pure, dans le cas de l'apprentissage profond. Nous chercherons à caractériser le prix de la pureté, à la fois dans le cas des attaques adversariales et dans les GANs.

3. **Convergence vers l'équilibre mixte approché.**

   Nous étudierons comment transposer les algorithmes d'apprentissage de stratégies mixtes et corrélés (comme le Fictitious Play, l'algorithme Follow the Regularized Leader [2]) au cas des réseaux profonds. Remarquons que les jeux liés aux GANs et aux attaques adversariales sont des jeux dans des domaines continus et en très grande dimension. Résoudre ces jeux de façon exacte est donc particulièrement difficile. Nous définirons les conditions sous lesquelles les algorithmes déjà existant peuvent s'appliquer à des jeux de cette dimension, et nous élaborerons de nouveaux algorithmes dans le but de converger rapidement vers un équilibre approché.

4. **Implémenter et appliquer notre modèle avec les algorithmes les plus performants sur des données réelles.**

   Enfin, il sera question d'appliquer les algorithmes élaborés précédemment pour valider notre théorie. Nous testerons notre methode sur plusieurs type de données. D'abord sur les bases de données connues comme MNIST, puis sur des données plus complexes : des images, des éléctrocardiogrammes etc..

References

[1] Ian J. Goodfellow and Jean Pouget-Abadie and Mehdi Mirza and Bing Xu and David Warde-Farley and Sherjil Ozair and Aaron Courville and Yoshua Bengio (2014) *Generative Adversarial Networks.* Neural Information Processing Systems.

[2] S. Shalev-Shwartz (2012) *Online learning and online convex optimization*, pages 107–194. Foundations and Trends in Machine Learning

[3] Hao Ge, Yin Xia, Xu Chen, Randall Berry, Ying Wu (2018) *Fictitious GAN: Training GANs with Historical Models.* ECCV.

[4] Papernot et al. (2016) *The limitations of deep learning in adversarial settings.* IEEE European Symposium on Security & Privacy