Ecole Graduée 631 MADIS

Sujet de thèse en Mathématique proposé en 2025

Titre : Des epsilon-réseaux aux entropies métriques

Directeur de thèse : Nicolas Wicker
E-mail : nicolas.wicker@univ-lille.fr

Co-directeur de thèse : Yann Guermeur
E-mail : yann.guermeur@loria.fr

Laboratoire : Paul Painlevé

Equipe : proba/stats

Descriptif : Le concept d'epsilon-réseau, introduit dans les années trente, a pris un rôle central en géométrie computationnelle avec l'article fondateur de Kolmogorov et Tihomirov (1961). Dans un espace métrique, un epsilon-réseau d'un domaine est un ensemble de points tel que tout point du domaine est situé à une distance strictement inférieure à epsilon d'un point de cet ensemble. Lorsqu'un domaine possède un epsilon-réseau de cardinal fini, alors son nombre de couverture est le plus petit cardinal de ses epsilon-réseaux. On nomme entropie métrique de logarithme de base deux de ce cardinal.

Au-delà de la géométrie, les entropies métriques jouent également un rôle central en théorie statistique de l'apprentissage (Vapnik, 1998). Lorsqu'elles sont calculées pour des familles de fonctions associées à un modèle de l'inférence empirique, que ce soit en discrimination ou en régression, leur comportement caractérise non seulement la consistance du principe inférentiel, mais encore la vitesse de convergence de l'erreur en apprentissage vers l'erreur en généralisation. Deux illustrations célèbres sont fournies par les machines à vecteurs support (Blanchard et al., 2008) et les réseaux de neurones (Bartlett et al., 2017). Dans le cas général, les entropies métriques sont majorées en fonction de dimensions combinatoires à facteur d'échelle (Kearns et R.E. Schapire, 1994 ; Guermeur, 2007) au moyen de « résultats combinatoires » aussi connus sous le nom de « lemmes de Sauer généralisés ». La majoration de ces dimensions fournit ensuite l'intervalle de confiance du risque garanti (majorant de l'erreur en généralisation).

Ce travail de thèse se compose de deux parties complémentaires. La première porte sur le calcul d'epsilon-réseaux de cardinalité minimale pour des ensembles finis de points. Ce problème est NP-difficile. Deux communautés ont développé des algorithmes fournissant des solutions approchées : celle de la classification (non supervisée) (Bien et Tibshirani, 2012 ; Moniot et al., 2022) et celle de la théorie des graphes (Li et al., 2020). Un premier objectif est d'effectuer une analyse synthétique de l'état de l'art, établissant le lien entre performance et complexité. Cette contribution initiale devrait donner naissance à de nouveaux algorithmes capables en particulier d'opérer dans des espaces non hilbertiens.

La seconde partie du travail de thèse relève de la discrimination à catégories multiples. Elle porte sur la majoration des entropies métriques des familles de fonctions associées aux systèmes discriminants à marge (réseaux de neurones, machines à noyau, forêts aléatoires...). Elle se décomposera suivant deux axes. Le premier consiste à améliorer les résultats combinatoires disponibles pour la principale dimension combinatoire dédiée à ces classifieurs : la dimension de Natarajan à marge (Guermeur, 2025). Le second est la majoration de

cette dimension pour les principaux systèmes discriminants de la littérature. Un intérêt tout particulier sera porté à la machine à noyau isotrope (Guermeur et Wicker, 2025) développée dans l'équipe. Pour cette machine, on peut espérer contrôler la capacité en fonction de l'isotropie des données (Ghorbani et al., 2020).

Bibliographie

P.L. Bartlett, D.J. Foster et M. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NIPS 30*, 2017.

J. Bien et R. Tibshirani. Hierarchical Clustering with prototypes via minimax linkage. J*ournal of the American Statistical Association*, 106 :1075-1084, 2012.

G. Blanchard, O. Bousquet et P. Massart. Statistical Performance of Support Vector Machines. *The Annals of Statistics*, 36(2) :489-531.

B. Ghorbani, S. Mei, T. Misiakiewicz, et A. Montanari. When do neural networks outperform kernel methods? In *NeurIPS 34*, 2020.

Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research,* 8:2551-2594, 2007.

Y. Guermeur. Sharper bounds on the metric entropies of margin classifiers. (soumis).

Y. Guermeur et N. Wicker. Isotropic kernel machine. (soumis).

M.J. Kearns et R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

A.N. Kolmogorov et V.M. Tihomirov. Epsilon-entropy and epsilon-capacity of sets in functional spaces. *American Mathematical Society Translations, series 2*, 17:277-364, 1961.

J. Li, R. Potru et F. Shahrokhi. A performance study of some approximation algorithms for computing a small dominating set in a graph. *Algorithms*, 13, 339, 2020.

A. Moniot, I. Chauvot de Beauchêne et Y. Guermeur. Inferring epsilon-nets of finite sets in a RKHS. In *WSOM+ 22*, 2022.

V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.