PAPER TITLE

0.1. **Comparaison RF-GBDT.**

Les forêts aléatoires et le *gradient boosting* paraissent très similaires au premier abord: il s'agit de deux approches ensemblistes, qui construisent des modèles très prédictifs performants en combinant un grand nombre d'arbres de décision. Mais en réalité, ces deux approches présentent plusieurs différences fondamentales:

- Les deux approches reposent sur des fondements théoriques différents: la loi des grands nombres pour les forêts aléatoires, la théorie de l'apprentissage statistique pour le *boosting*.
- Les arbres n'ont pas le même statut dans les deux approches. Dans une forêt aléatoire, les arbres sont entraînés indépendamment les uns des autres et constituent chacun un modèle à part entière, qui peut être utilisé, représenté et interprété isolément. Dans un modèle de *boosting*, les arbres sont entraînés séquentiellement, ce qui implique que chaque arbre n'a pas de sens indépendamment de l'ensemble des arbres qui l'ont précédé dans l'entraînement.
- Les points d'attention dans l'entraînement ne sont pas les mêmes: arbitrage puissance-corrélation dans la RF, arbitrage puissance-overfitting dans le *boosting*.
- *overfitting*: borne théorique à l'*overfitting* dans les RF, contre pas de borne dans le *boosting*. Deux conséquences: 1/ lutter contre l'overfitting est essentiel dans l'usage du *boosting*; 2/ le *boosting* est plus sensible au bruit et aux erreurs sur y que la RF.
- Conditions d'utilisation: la RF peut être utilisée en OOB, pas le *boosting*.
- Complexité d'usage: peu d'hyperparamètres dans les RF, contre un grand nombre dans le *boosting*.

References