# L'apprentissage statistique à grande échelle

Léon Bottou[1], Olivier Bousquet[2]

[1] NEC Labs America, Princeton, USA
[2] Google, Zurich, Suisse

**Résumé** Depuis une dizaine d'années, la taille des données croit plus vite que la puissance des processeurs. Lorsque les données disponibles sont pratiquement infinies, c'est le temps de calcul qui limite les possibilités de l'apprentissage statistique. Ce document montre que ce changement d'échelle nous conduit vers un compromis qualitativement différent dont les conséquences ne sont pas évidentes. En particulier, bien que la descente de gradient stochastique soit un algorithme d'optimisation médiocre, on montrera, en théorie et en pratique, que sa performance est excellente pour l'apprentissage statistique à grande échelle.

## 1 Introduction

La théorie de l'apprentissage statistique prend rarement en compte le coût des algorithmes d'apprentissage. Vapnik [1] ne s'y intéresse pas. Valiant [2] exclue les algorithmes d'apprentissage dont le coût croit exponentiellement. Cependant, malgré de nombreux progrès sur les aspects statistiques [3, 4], peu de résultats concerne la complexité des algorithmes d'apprentissage (e.g., [5].)

Ce document développe une idée simple : une optimisation approximative est souvent suffisante pour les besoins de l'apprentissage. La première partie reprend la décomposition de l'erreur de prévision proposée dans [6] dans laquelle un terme supplémentaire qui décrit les conséquences de l'optimisation approximative. Dans le cas de l'apprentissage à petite échelle, cette décomposition décrit le compromis habituel entre approximation et estimation. Dans le cas de l'apprentissage à grande échelle, elle décrit une situation plus complexe qui dépend en particulier du coût de calcul associé à l'algorithme d'apprentissage. La seconde partie explore les propriétés asymptotiques de l'apprentissage à grande échelle lorsque l'on utilise diverses méthodes d'optimisation. Ces résultats montrent clairement que le meilleur algorithme d'optimisation n'est pas nécessairement le meilleur algorithme d'apprentissage. Finalement, cette analyse est confirmée par quelques comparaisons expérimentales.

## 2 Optimisation approximative

Comme [7, 1], considérons un espace de paires entrées-sorties $(x, y) \in \mathcal{X} \times \mathcal{Y}$ équippé d'une loi jointe de probabilité $P(x, y)$. La loi conditionelle $P(y|x)$ représente la relation inconnue qui lie entrées et sorties. Une fonction de perte $\ell(\hat{y}, y)$ mesure l'écart entre la sortie prédite $\hat{y}$ et la sortie observée $y$. Notre objectif est la fonction $f^*$ qui minimise le *risque moyen*

$$E(f) = \int \ell(f(x), y)\, dP(x, y) = \mathbb{E}\left[\ell(f(x), y)\right],$$