Adaptativity of Stochastic Gradient Descent
Aymeric Dieuleveut  F. Bach, Non parametric stochastic approximation with large step sizes, in the Annals of Statistics

Setting : random-design least-squares regression problem in a RKHS framework.
Risk : for $g : \mathcal{X} \rightarrow \mathbb{R}$

$$\varepsilon(g) := \mathbb{E}_{\rho} (g(X) - Y)^2.$$ 

We thus want to minimize prediction error.

Regression function : $g_{\rho}(X) = \mathbb{E}[Y|X]$ minimises $\varepsilon$ on $L^2_{\rho X}$.

We build a sequence $(g_k)$ of estimators in an RKHS $\mathcal{H}$.

Why considering RKHS ?

- hypothesis space for non parametric regression,
- high dimensional problem ($d >> n$) analysis framework,
- natural analysis when mapping data in feature space via a p.d. kernel.
Regularity assumptions

Algorithm (Stochastic approximation)

Simple one pass stochastic gradient descent with constant step sizes and averaging.

Difficulty of the problem

- Let \( \Sigma = \mathbb{E}[K_x K_x^t] \) be the covariance operator. We assume that
  \( \text{tr}(\Sigma^{1/\alpha}) < \infty \)
- We assume \( g_\rho \in \Sigma^r(L^2_{\rho X}) \).

(\( \alpha, r \)) encode the difficulty of the problem.
Theorem (Non parametric regression)

Under a suitable choice of the learning rate, we get the optimal rate of convergence for non parametric regression.

Theorem (Adaptativity in Euclidean spaces)

If $\mathcal{H}$ is a $d$-dimensional Euclidean space:

$$\mathbb{E} [\varepsilon (\tilde{g}_n) - \varepsilon (g_\rho)] \leq \min_{1 \leq \alpha, -\frac{1}{2} \leq q \leq \frac{1}{2}} \frac{16 \sigma^2 \text{tr}(\varSigma^{1/\alpha})(\gamma n)^{1/\alpha}}{n} + 8 \frac{||T^{-q}\theta_{\mathcal{H}}||^2_{\mathcal{H}}}{(n\gamma)^{2q+1}}.$$

$SGD$ is adaptative to the regularity of the objective function and to the decay of the spectrum of the covariance matrix.

$\Leftarrow$ explains behaviour for $d \gg n$. 