

Online Learning Algorithms as Stochastic Approximations of the Regularization Path

PIERRE TARRÈS

(joint work with Yuan Yao)

Consider the following classical problem of learning from examples: given a sequence of i.i.d. random samples $(z_t = (x_t, y_t))_{t \in \mathbb{N}}$ drawn from a probability measure ρ on $X \times Y$, one seeks to approximate the *regression function*

$$f_\rho(x) := \int_Y y d\rho_{Y|x},$$

i.e., the conditional expectation of y given x .

We study here *online learning algorithms*, which are recursive, contrary to *batch learning algorithms* which process the data once and for all at some fixed time m . We show [7], using stochastic approximation techniques, how their convergence rates can match the batch learning ones.

The quality of the estimate one can obtain depends on the regularity of f_ρ , measured through a Mercer kernel $K : X \times X \rightarrow \mathbb{R}$ (continuous, symmetric and positive semidefinite). The Reproducing Kernel Hilbert Space (RKHS) \mathcal{H}_K is defined as the closure of the linear span of the set of functions $\{K_x := K(x, \cdot), x \in X\}$, with the inner product, denoted as $\langle \cdot, \cdot \rangle_K$, satisfying $\langle K_x, K_y \rangle_K = K(x, y)$.

Recall the reproducing property $\langle K_x, f \rangle = f(x)$, for all $x \in X, f \in \mathcal{H}_K$, which implies in particular that $\|f\|_\infty \leq \kappa \|f\|_K$, where $\kappa := \sup_{x \in X} \sqrt{K(x, x)}$.

We analyze *online* algorithms of the type

$$f_t = f_{t-1} - \gamma_t [(f_{t-1}(x_t) - y_t)K_{x_t} + \lambda_t f_{t-1}], \quad \text{for some } f_0 \in \mathcal{H}_K, \text{ e.g. } f_0 = 0,$$

with gain sequences $(\lambda_t)_{t \in \mathbb{N}}$ and $(\gamma_t)_{t \in \mathbb{N}}$ taking values in $\mathbb{R}_+ \setminus \{0\}$, originally introduced by Smale and Yao in [5], and further studied by Yao in [8]. The recursion can be interpreted as a stochastic gradient descent

$$f_t = f_{t-1} - \text{grad } V_{z_t}^{\lambda_t}(f_{t-1}),$$

where

$$V_z^\lambda(f) := \frac{1}{2} [(f(x) - y)^2 + \lambda \|f\|_K^2]$$

for all $f \in \mathcal{H}_K, z \in X \times Y$ and $\lambda \in \mathbb{R}_+$. One of the advantages of such algorithms is their computational complexity, which is quadratic in time in the worst case, and can be linear at the cost of a large memory allocation. In comparison, the batch learning Tikhonov regularization scheme typically involves the inverse of a matrix, which is $O(t^3)$ in the worst case.

We optimize the choice of $(\lambda_t)_{t \in \mathbb{N}}$ and $(\gamma_t)_{t \in \mathbb{N}}$, as a function of the regularity of f_ρ . More precisely, let ρ_X be the induced marginal probability measure from ρ on X , and let $L_K : \mathcal{L}^2(\rho_X) \rightarrow \mathcal{L}^2(\rho_X)$ be the self-adjoint operator defined by

$$L_K(f)(x) = \int_X K(x, y) f(y) d\rho_X(y) = \langle K_x, f \rangle_{\mathcal{L}^2(\rho_X)}, \quad x \in X,$$

which is positive and compact, so that we can define (through any orthonormal system), the operators $L_K^r : \mathcal{L}^2(\rho_X) \longrightarrow \mathcal{L}^2(\rho_X)$ for all $r \in \mathbb{R}_+$.

Assume that f_ρ lies in the image of L_K^r . We show that, if we choose $f_0 := 0$, and

$$(1) \quad \gamma_t := a(t + t_0)^{-\frac{2r}{2r+1}}, \quad \lambda_t := b(t + t_0)^{-\frac{1}{2r+1}},$$

for some $t_0 := \text{Cst}(\kappa)$, $a, b := \text{Cst}(M_\rho, \|L_K^{-r} f_\rho\|_K)$ then, with confidence $1 - \delta$,

$$\|f_t - f_\rho\|_K \leq \text{Cst}(\kappa, M_\rho, \|L_K^{-r} f_\rho\|_{\mathcal{L}^2(\rho_X)}) \left(\log \frac{2}{\delta} \right) t^{-\frac{2r-1}{4r+2}},$$

and

$$\|f_t - f_\rho\|_{\mathcal{L}^2(\rho_X)} \leq \text{Cst}(\kappa, M_\rho, \|L_K^{-r} f_\rho\|_{\mathcal{L}^2(\rho_X)}) \left(\log \frac{2}{\delta} \right)^2 t^{-\frac{r}{2r+1}}.$$

The choice $a = b := 1$ yields the same result, at the expense, however, of the constants involved.

The exponent in t in the \mathcal{H}_K -norm rate is the same as the best known one in batch learning, obtained by Smale and Zhou [6], and the mean square distance convergence rate is optimal in the sense that it reaches the minimax and individual lower rates (see for instance Caponnetto and de Vito [2]).

The choice of these gain sequences in (1) is derived from the analysis of the algorithm as a stochastic approximation of a Tikhonov regularization path converging to the regression function.

In the talk we explain some previous results on the convergence rates of stochastic algorithms, in particular the “1/2-phase transition”, which also plays an important rôle in the Pólya urn model (see for instance Athreya and Karlin [1] or, more recently, Pouyanne [4]). We show how these finite-dimensional techniques can be extended to the infinite-dimensional online algorithm considered here, using on the one hand some martingale and reverse-martingale expansions, and on the other hand probabilistic exponential inequalities on Banach spaces provided by Pinelis [3].

REFERENCES

- [1] K. B. Athreya and S. Karlin, *Embedding of urn schemes into continuous time Markov branching processes and related limit theorems*, Ann. Math. Statist. **39** (1968), 1801–1817.
- [2] A. Caponnetto and E. De Vito, *Optimal rates for regularized least squares algorithm*, Found. Comput. Math. **7**, no. 3 (2007), 331–368.
- [3] I. Pinelis, *Optimum bounds for the distributions of martingales in Banach spaces*, Ann. Prob. **22**, no. 4 (1994), 1679–1706.
- [4] N. Pouyanne, *An algebraic approach to Pólya processes*, Ann. Inst. H. Poincaré, **44**, no. 2 (2008), 293–323.
- [5] S. Smale and Y. Yao, *Online learning algorithms*, Found. Comput. Math. **6**, no. 2 (2006), 145–170.
- [6] S. Smale D.-X. Zhou, *Learning theory estimates via integral operators and their approximations*, Constr. Approx. **26**, no. 2 (2007), 153–172.
- [7] P. Tarrès and Y. Yao, *Online learning algorithms as stochastic approximations of the regularization path*, Preprint, 2006.