Online Learning with Markov Sampling (a presentation)*

David Barrera

Abstract

In this note we provide an interpretation of the online algorithm by Smale and Zhou in [3] in terms of Stochastic Gradient Descent techniques. In particular we will see that, modulo some heuristics, the algorithm might be consistent for general sampling sequences that are convergent in distribution and satisfy a pointwise ergodic theorem.

1 Introduction

We depart from a random vector

$$D_n = ((X_k, Y_k))_{k=1}^n : \Omega \to (S \times \mathbb{R})^n \tag{1}$$

where $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space. Thus S is the "state" space ($X_k \in S$ for all k) and $\mathbb{R}$ is the "response" space ($Y_k \in \mathbb{R}$ for all k). Throughout these notes, we will not assume that the random vectors $(X_k, Y_k)_k$ are i.i.d.

1.1 Statement of the Problem in terms of Loss Functions

Suppose that we are given a family $\mathcal{F}$ of Borel measurable functions $f : S \to \mathbb{R}$ and a "loss function"

$$L_n : (\mathbb{R} \times \mathbb{R})^n \to [0, \infty). \tag{2}$$

Definition 1 (The Problem of Regression)**.** *The* ***problem of regression*** *for the loss function L_n, the distribution D_n, and the family $\mathcal{F}$ as above, is that of finding (if possible) an element*

$$f^*(\mathcal{F}, L_n, D_n) \in \arg\min_{f \in \mathcal{F}} EL_n((f(X_k), Y_k)_{k=1}^n). \tag{3}$$

Thus, if L_n measures in some sense the distance between the components of a vector $((r_k, y_k)_{k=1}^n) \in (\mathbb{R} \times \mathbb{R})^n$, then $f^*(\mathcal{F}, L_n, D_n)$ in (3) represents the best predictor (on average) of Y as a function of X according to the distribution of D_n.

Remark 1. To bypass the problem of the existence of $f^*(\mathcal{F}, L_n, D_n)$ we can change Definition 1 to the requirement of finding, for every given $\epsilon > 0$, an element $f_\epsilon^* = f^*(\epsilon, \mathcal{F}, L_n, D_n)$ with

$$EL_n((f_\epsilon^*(X_k), Y_k)_{k=1}^n) - \inf_{f \in \mathcal{F}} EL_n((f(X_k), Y_k)_{k=1}^n) < \epsilon, \tag{4}$$

a problem of which (3) is always a solution (for all $\epsilon > 0$) when it exists.

*These notes are the memory of a talk given on May 11th, 2017 at the École Polytechnique as part of the activities of the reading group on Machine Learning. The purpose of the talk was to introduce the results and methods in [3].

Example (The Empirical L^2 Norm)**.** If

$$L_n((r_k, y_k)_{k=1}^2) = \frac{1}{n}\sum_{k=1}^{n} |r_k - y_k|^2,$$

then (3) corresponds to finding a solution to

$$f^*(\mathcal{F}, D_n) := \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} E|f(X_k) - Y_k|^2. \tag{5}$$

As we see immediately, the average at the right-hand side of (5) is nothing but $E|f(X) - Y|^2$ in the case in which $(X_k, Y_k)_k \sim (X, Y)$ for some random vector (X, Y) (for instance in the i.i.d. case).

1.2 Notation for Regression

Definition 2 (Regression Objects)**.** *We will call the problem of finding (5) the (nonasymptotic)* ***least-squares regression problem (within $\mathcal{F}$) for the distribution of*** *D_n, we will call $f^*(\mathcal{F}, D_n)$ a D_n-**regressor of** y **as a function of** x **within the family** $\mathcal{F}$ and, given $\hat{D}_n := ((x_k, y_k)_{k=1}^n) \in (S\times\mathbb{R})^n$ (typically "a realization" of D_n), we will call an element*

$$\hat{f}^*(\mathcal{F}, \hat{D}_n) \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} |f(x_k) - y_k|^2, \tag{6}$$

a ***solution to the least-squares regression problem associated to $\mathcal{F}$ and $\hat{D}_n$*** *(note that no reference to a distribution is made here).*

With this notation $\omega \mapsto \hat{f}^*(\mathcal{F}, D_n(\omega))$ is a random variable[1]. In this language, the problem of ***strong consistency*** is that of proving that one has the convergence

$$\lim_{n\to\infty} |f^*(\mathcal{F}, D_n) - \hat{f}^*(\mathcal{F}, D_n(\omega))| = 0, \tag{7}$$

for $\mathbb{P}-$a.e. ω.

As we know, the problem of strong consistency is not trivial even in the i.i.d. case, its solution requiring additional hypotheses such as restrictions on the complexity of $\mathcal{F}$ or different approaches to the empirical approximation of $f^*(\mathcal{F}, D_n)$, such as "complexity regularization". This last setting is important for what follows:

Definition 3 (Complexity Regularization)**.** *Let $\Lambda := (\lambda_k)_k$ be a sequence of nonnegative numbers. In the setting of Definition 2, and assuming that $\mathcal{F}$ has the structure of (a subset of a) normed vector space $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$, we will call*

$$f^*(\mathcal{F}, D_n, \Lambda) \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} (E|f(X_k) - Y_k|^2 + \lambda_k||f||_{\mathcal{F}}^2) \tag{8}$$

a $\Lambda-$***regularized*** D_n-***regressor of*** y ***as a function of*** x ***within*** $\mathcal{F}$***.***

Note that, with this definition, a D_n-regressor of y as a function of x within $\mathcal{F}$ (Definition 2) corresponds to the choice of $\Lambda = (0, 0, \dots)$, and that if $\lambda_k = \lambda$ and $(X_k, Y_k) \sim (X, Y)$ for all $k \in \mathbb{N}$ then (8) is nothing but the "penalized" least-squares regression problem

$$\arg\min_{f\in\mathcal{F}} (E|f(X) - Y|^2 + \lambda||f||_{\mathcal{F}}^2). \tag{9}$$

[1] In this note, we will skip measurability problems.

1.3 An Observation

The following observation constitutes, in our presentation, the "intuitive principle" behind the convergence of the algorithm proposed in [3]: suppose that $\cdots \subset D_n \subset D_{n+1} \subset \ldots$ is given according to (1), i.e., that for all n, $D_{n+1} = (D_n, (X_{n+1}, Y_{n+1}))$, that we are given a corresponding sequence of functions $(f^*(\mathcal{F}, D_n, \Lambda))_{n\in\mathbb{N}}$ according to (8) and that, in addition, $\lambda_n \to_n 0$ or, more generally, that

$$\lim_n \frac{1}{n}\sum_{k=1}^{n} \lambda_k = 0. \tag{10}$$

Suppose also that, for all $f \in \mathcal{F}$, one has the existence of the limit

$$\varphi_\omega(f) := \lim_n \frac{1}{n}\sum_{k=1}^{n} |f(X_k(\omega)) - Y_k(\omega)|^2 \tag{11}$$

for $\mathbb{P}-$a.e. ω.

Since the map $\varphi_{\omega,n}$ given by

$$f \mapsto \frac{1}{n}\sum_{k=1}^{n}(|f(X_k(\omega)) - Y_k(\omega)|^2 + \lambda_k||f||^2_{\mathcal{F}})$$

is convex for a fixed $\omega \in \Omega$ and satisfies

$$\lim_n \varphi_{\omega,n}(f) = \varphi_\omega(f),$$

$\mathbb{P}-$a.s., then the function φ_ω is (as well) convex for every fixed ω where the limit (11) exists, and if f^*_ω is a limit point (with respect to $||\cdot||_{\mathcal{F}}$) of

$$\hat{f}^*(\mathcal{F}, D_n(\omega), \Lambda))_n \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} |f(X_k(\omega)) - Y_k(\omega)|^2$$

then, since

$$\hat{f}^*(\mathcal{F}, D_n(\omega), \Lambda) \in \arg\min_{f\in\mathcal{F}} \varphi_{\omega,n}(f)$$

we can expect that[2]

$$f^*_\omega \in \arg\min_{f\in\mathcal{F}} \varphi_\omega(f) \tag{12}$$

provided that such arg min indeed exists.

2 Stochastic Gradient Descent

We review now the basic idea of the algorithm of (iterative) Stochastic Gradient Descent, for details see for instance [1] (or Wikipedia).

[2] If $\psi_n : V \to \mathbb{R}$ is a family of continuous functions defined on some normed vector space V and converging pointwise to ψ, then any convergent sequence of minimizers of ψ_n converges to a minimizer of ψ *provided that* $\psi_n(v_n) - \psi_n(v) \to_n 0$ whenever $v_n \to_n v$, in particular if $(\psi_n)_n$ is a uniformly Lipschitz family with bounded Lipschitz constants (it might be instructive to think about the possible obstructions to these hypothesis and/or their conclusion in the case under consideration).

This can be seen considering the counter-reciprocal argument: given any point $v \in V$ that is *not* a minimizer of ψ, so that for some $v' \in V$ and some $\epsilon > 0$

$$\psi(v) > \psi(v') + 4\epsilon,$$

then for any sequence of points v_n with $v_n \to_n v$ one has, for $n \geq N$, that

$$\psi_n(v') + \epsilon < \psi(v') + 2\epsilon < \psi(v) - 2\epsilon < \psi_n(v) - \epsilon < \psi_n(v_n),$$

so that $(v_n)_n$ cannot be a sequence of minimizers of $(\psi_n)_n$.

Stochastic Gradient Descent (Sketch). *Suppose that we are given a "good" vector space W, a function $\ell : W \times Z \to \mathbb{R}$ (convex for every fixed $z \in Z$), and a set of points $(z_1, \dots, z_n) \in Z$. Then in order to approximate*

$$w^* \in \arg\min_{w \in W} \frac{1}{n} \sum_{k=1}^{n} \ell(w, z_k)$$

we can

1. *Select (conveniently) an initial point $w_0 \in W$.*
2. *Given $0 \leq k < n$ and $w_0, \dots, w_k \in W$, choose a "learning step" p_k and do*

$$w_{k+1} = w_k - p_k D_w \ell(w_k, z_k). \tag{13}$$

3. *For $k = n$, repeat the steps 1. and 2. starting with $w'_0 = w_n$ and $(z'_1, \dots, z'_n) = (z_{\sigma(1)}, \dots, z_{\sigma(n)})$ for some permutation σ of the index set $(1, \dots, n)$.*

We actually will not use the step 3. of the algorithm above to deduce the generic form of our learning procedure, which we proceed now to explain.

3 The Algorithm and its Rate (under Assumptions)

Suppose that we want to approximate (8) under the hypothesis that $\mathcal{F}$ is the ***reproducing kernel Hilbert space*** $\mathcal{H}_K$ associated to a Mercer Kernel (see [2])

$$K : S \times S \to \mathbb{R}.$$

This is, $\mathcal{H}_K$ is the completion of the vector space generated by the elements in the sequence of functions $(K(x, \cdot))_{x \in S} =: (K_x)_{x \in S}$ with respect to the inner product

$$< K_{x_1}, K_{x_1} >_K =: K(x_1, x_2).$$

For the purpose of approximating (8) in this space, we depart from an observation $D_n(\omega) = ((X_k(\omega), Y_k(\omega)))_{k=1}^n$ of D_n and we use the empirical estimator

$$\hat{f}^*(K, D_n(\omega), \Lambda) \in \arg\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{k=1}^{n} (|f(X_k(\omega)) - Y_k(\omega)|^2 + \lambda_k ||f||^2_{\mathcal{F}}) \tag{14}$$

Our goal now is to approximate $\hat{f}^*(K, D_n(\omega), \Lambda)$ in an online way. To do so, we use the algorithm of Stochastic Gradient Descent with the loss function $\ell : \mathcal{H}_k \times (S \times \mathbb{R} \times [0, \infty)) \to [0, \infty)$ defined by

$$\ell(f, (x, y, \lambda)) := (f(x) - y)^2 + \lambda ||f||^2_K = (K_x - y\mathbf{1})^2 f + \lambda < f, f >_K \tag{15}$$

where $\mathbf{1}$ is the constant function $f \mapsto 1$ defined on $\mathcal{H}_K$.

Exercise: *Prove that the Fréchet derivative of ℓ with respect to f is*

$$D_f \ell(f_0, (x, y, \lambda)) = 2(f_0(x) - y) < K_x, \cdot >_K + 2\lambda < f_0, \cdot >_K, \tag{16}$$

and therefore that we can think of $D_f \ell(f_0, (x, y, \lambda))$, via the Riesz Representation Theorem, as the element

$$D_f \ell(f_0, (x, y, \lambda)) = 2(f_0(x) - y) K_x + 2\lambda f_0 \tag{17}$$

of $\mathcal{H}_K$.

Together, (17) and (13) give rise to the following Stochastic Gradient Descent procedure:

To approximate (14):

1. *Select an initial function $f_{\omega,0} \in \mathcal{H}_K$, say $f_0(\omega) = 0$.*

2. *Given $f_{\omega,0}, \ldots, f_{\omega,k}$, pick a step size p_k and let*

$$f_{\omega,k+1} := f_{\omega,k} - p_k((f_{\omega,k}(X_k(\omega)) - Y_k(\omega))K_{X_k(\omega)} + \lambda_k f_{\omega,k}). \tag{18}$$

Now, since we are in an "online setting" in which the information $(X_k(\omega), Y_k(\omega))_k$ flows without any a priori bound n in k, and since we expect (from the stochastic gradient descent) that

$$\lim_n |f_{\omega,n} - \hat{f}^*(K, D_n(\omega), \Lambda)| = 0,$$

$\mathbb{P}-$a.s., then and application of the observation in Section 1.3 authorizes us to expect that

$$\lim_n f_{\omega,n} = \lim_n \hat{f}^*(K, D_n(\omega), \Lambda) = \arg\min_{f \in \mathcal{H}_K} \lim_n \frac{1}{n} \sum_{k=1}^n |f(X_k(\omega)) - Y_k(\omega)|^2,$$

provided that there exists

$$\lim_n \frac{1}{n} \sum_{k=1}^n |f(X_k(\omega)) - Y_k(\omega)|^2 \tag{19}$$

for $\mathbb{P}-$a.e. ω.

And what is a natural candidate for (19)? If we assume that $(X_k, Y_k) \Rightarrow_k (X, Y)$ under "ergodic conditions"[3] ("$\Rightarrow$" denotes convergence in distribution), the natural guess is to have

$$\lim_n \frac{1}{n} \sum_{k=1}^n |f(X_k(\omega)) - Y_k(\omega)|^2 = E|f(X) - Y|^2,$$

$\mathbb{P}-$a.s. This bring us finally to the main result in [3]:

Theorem (Simplified Version of [3], Theorem 1)**.** *Under the assumptions 1.-5. below and their notation, let $p, \lambda, \beta, \theta > 0$ be such that $0 \leq \beta \leq 1-\theta$, and define the sequence of random functions $\mathbf{f} = (f_k)$ according to (18)[4] for $p_k = pk^{-\theta}$ and $\lambda_k = \lambda k^{-\beta}$. Then we have the nonasymptotic bounds*

$$E[||f_{k+1} - f_\infty||_K | D_k] \leq \begin{cases} Ck^{-\min\{\beta(r-1/2), (\theta-\beta)/2\}} & , \quad 0 < \beta < 1-\theta \\ Ck^{-\min\{\beta(r-1/2), (\theta-\beta)/2\}} \log(k+1) & , \quad \beta = 1-\theta \\ Ck^{-\theta/2} & , \quad \beta = 0 \end{cases} \tag{20}$$

where $C = C(\alpha, \beta, \lambda, \theta, \kappa_{2,\delta}, ||K||_{C^\delta(S\times S)}, r)$. If we assume that $\alpha = 1$, we still have the (nonconvergent) bound Ck^θ for the left-hand side of (20).

Assumptions:

We are given a "target" random vector $(X_\infty, Y_\infty) : \Omega \to S \times \mathbb{R}$ under the assumptions that

1. *(Compactness of S)* The state space of X_k, S, is a compact metric space with metric d (in particular S is separable).

2. *(Exponential Convergence)* If ρ_{X_k} denotes the marginal distributions of X_k, then

$$||\rho_{X_k} - \rho_{X_\infty}||_{(C^{0,\delta}(S))^*} \leq C\alpha^k \tag{21}$$

for some $\alpha \in (0,1)$, where $C^{0,\delta}(S)$ is the space of $\delta-$Hölder continous functions on S ($\delta \geq 0$).

[3] We are deliberately ambiguous here.

[4] Thus $\mathbf{f}(\omega) := (f_{\omega,k})_k$.

3. *(Consistent Predictability)* There exists a function f_∞ with the property that and for all $k \in \mathbb{N} \cup \{\infty\}$,

$$E[Y_k|X_k] = f_\infty \circ X_k,$$

$\mathbb{P}-$a.s.

4. *(Lipschizity of the Mercer Kernel)* The Kernel K satisfies the following $\delta-$Lipschitz (product) condition

$$|(K(x_1, x_1') - K(x_2, x_1')) - (K(x_1, x_2') - K(x_2, x_2'))| \leq \kappa_{2,\delta}(d(x_1, x_2))^\delta (d(x_1', x_2'))^\delta. \tag{22}$$

5. *(Regularity of f_∞).* For some $r \in (1/2, 3/2)$,[5] the regression function f_∞ from the item 3. can be chosen in $Q^r_{K,\rho_{X_\infty}} L^2_{\rho_{X_\infty}}$, where

$$Q_{K,\mu} f(\cdot) = \int_S K(\cdot, y) f(y) d\mu(y).$$

This implies in particular that f_∞ can be seen as an element of $\mathcal{H}_K$ via the Hilbert space isomorphism $Q^{1/2}_{K,\rho_{X_\infty}}$ between $L^2_{\rho_{X_\infty}}$ and $\mathcal{H}_K$, see [2], and therefore that

$$f_\infty \in \arg\min_{f \in \mathcal{H}_k} E|f(X_\infty) - Y_\infty|^2$$

which is consistent with what we expect from the arguments above.

References

[1] Bottou, L (1998). Online Algorithms and Stochastic Approximations. Online Learning and Neural Networks. *Cambridge University Press*

[2] Cucker, F and Smale, S. (2000). On the Mathematical Fundations of Learning. *Bull. of the AMS* **39**. Vol 1. Pp. 1–49.

[3] Smale, S. and Zhou, D-X. (2007). Online Learning with Markov Sampling. (As in the preprint avaliable at D-X Zhou's webpage).

[5] I do not see this restriction in the hypotheses of the paper, but one finds its necessity checking the proofs.