Adaptativity of Stochastic Gradient Descent

Aymeric Dieuleveut
Ecole Normale Superieure, Paris
aymeric.dieuleveut@ens.fr

Francis Bach
Ecole Normale Superieure, Paris
francis.bach@ens.fr

Abstract

We consider the random-design least-squares regression problem within the reproducing kernel Hilbert space (RKHS) framework. Given a stream of independent and identically distributed input/output data, we aim to learn a regression function within an RKHS $\mathcal{H}$, even if the optimal predictor (i.e., the conditional expectation) is not in $\mathcal{H}$. In a stochastic approximation framework where the estimator is updated after each observation, we show that the averaged unregularized least-mean-square algorithm (a form of stochastic gradient descent), given a sufficient large step-size, attains optimal rates of convergence for a variety of regimes for the smoothnesses of the optimal prediction function and the functions in $\mathcal{H}$. Our results apply as well in the usual finite-dimensional setting of parametric least-squares regression, showing adaptivity of our estimator to the spectral decay of the covariance matrix of the covariates.

1 Introduction

This abstract presents results described with more details in [1], which will soon appear in the Annals of Statistics[1] : we analyze non-parametric least-squares regression within the RKHS framework. RKHS provide an interesting tool to analyze high dimensional problem. Moreover it is also classical to use kernels spaces as an hypothesis space for non-parametric regression, approximating the regression function by a sequence of functions in the kernel space; or to use kernels to map non-vectorial data into a linear space, which allows to separate of the representation problem (designing good kernels) and the algorithmic/theoretical problems (given a kernel, how to design, run efficiently and analyse estimation algorithms).

We follow a stochastic approximation framework formulated directly in the RKHS, in which each observation is used only once and overfitting is avoided by making only a single pass through the data. Traditional online stochastic approximation algorithms, as introduced by Robbins and Monro [2], lead to stochastic gradient descent methods with step-sizes decreasing with the number of observations n, which are typically proportional to $n^{-\zeta}$, with ζ between $1/2$ and 1. Here, following [3] we show that using longer step-sizes with averaging also brings to Hilbert space settings needed for non parametric regression.

We characterize the convergence rate of averaged least-mean-squares and show how the proper set-up of the step-size leads to optimal convergence rates (as they were proved in [4]), extending results from finite-dimensional [3] to infinite-dimensional settings. The problem we solve here was stated as an open problem in [5, 6]. Moreover, our results apply as well in the usual finite-dimensional setting of parametric least-squares regression, showing adaptivity of our estimator to the spectral decay of the covariance matrix of the covariates.

2 Setting

Minimization problem. In this paper, we consider a general random design regression problem, where observations (x_i, y_i) are independent and identically distributed (i.i.d.) ran-

[1] and has never been presented to any workshop or machine learning conference.

dom variables in $\mathcal{X} \times \mathcal{Y}$ drawn from a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$. The set $\mathcal{X}$ is assumed to be a compact set with a full support measure; and we consider for simplicity $\mathcal{Y} = \mathbb{R}$. We measure the risk of a function $g : \mathcal{X} \to \mathbb{R}$, by the mean square error, that is, $\varepsilon(g) := \mathbb{E}_\rho\left[(g(X) - Y)^2\right]$. We denote by ρ_X the marginal law on the space $\mathcal{X}$. We may use the notations $\mathbb{E}_{\rho_X}[f(\cdot)]$ for $\int_{\mathcal{X}} f(x) d\rho_X(x)$. We denote by $\|\cdot\|_{L^2_{\rho_X}}$ the norm: $\|f\|^2_{L^2_{\rho_X}} = \int_{\mathcal{X}} |f(x)|^2 d\rho_X(x)$, over the space $L^2_{\rho_X}$ of squared integrable functions with respect to ρ_X. The function g that minimizes $\varepsilon(g)$ over $L^2_{\rho_X}$ is known to be the conditional expectation, that is, $g_\rho(X) = \mathbb{E}[Y|X]$. Note that we aim to minimize the *prediction error* of a function f. An important property of the prediction error is that the excess risk may be expressed as a squared distance to g_ρ : for any $f \in L^2_{\rho_X}, \varepsilon(f) - \varepsilon(g_\rho) = \|f - g_\rho\|^2_{L^2_{\rho_X}}$.

Hypothesis space. In this paper we consider formulations where our estimates lie in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with a continuous positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (K is a mercer Kernel). An RKHS satisfies :

1) $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a separable Hilbert space of functions: $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$.
2) $\mathcal{H}$ contains all functions $K_x : t \mapsto K(x, t)$, for all x in $\mathcal{X}$.
3) For any $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the reproducing property holds: $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$.

The reproducing property allows to treat non-parametric estimation in the same algebraic framework as parametric regression.

A key feature of our analysis is that we only considered $\|f - g_\rho\|^2_{L^2_{\rho_X}}$ as a measure of performance and do not consider convergences in stricter norms. This allows us to neither assume that g_ρ is in $\mathcal{H}$ nor that $\mathcal{H}$ is dense in $L^2_{\rho_X}$. We define $g_{\mathcal{H}} = \arg\min_{f \in \overline{\mathcal{H}}} \varepsilon(g)$, the best function over the closure $\overline{\mathcal{H}}$ of $\mathcal{H}$.

Moment assumptions. We make the following simple assumption regarding finiteness of moments: $R^2 := \sup_{x \in \mathcal{X}} K(x, x)$ and $\mathbb{E}[Y^2]$ are finite. Note that under these assumptions, any function in $\mathcal{H}$ is in $L^2_{\rho_X}$.

Covariance operator. We define the covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$, for the space $\mathcal{H}$ and probability distribution ρ_X, through $\forall (f, g) \in \mathcal{H}^2, \quad \langle f, \Sigma g \rangle_{\mathcal{H}} = \mathbb{E}[f(X) g(X)]$. Using the reproducing property, we have: $\Sigma = \mathbb{E}[K_X \otimes K_X]$, where for any elements $g, h \in \mathcal{H}$, we denote by $g \otimes h$ the operator from $\mathcal{H}$ to $\mathcal{H}$ defined as: $g \otimes h \ : f \mapsto \langle f, h \rangle_{\mathcal{H}} \ g$. The spectral properties of such an operator have appeared to be a key point to characterize the convergence rates of estimators [7, 8, 4].

In finite dimension, Σ is the usual (non-centered) covariance matrix. Moreover, it is possible to extend such an operator as an endomorphism T from $L^2_{\rho_X}$ to $L^2_{\rho_X}$, see [1] for details . Such an extension can be proved to be a Hilbert Schmidt operator, which allows to define the powers T^r, for $r \geqslant 0$, which will be used to quantify the regularity of the function $g_{\mathcal{H}}$.

Regularity. We make the following assumptions :

A1. We assume $\mathrm{tr}(T^{\frac{1}{\alpha}}) < s^{2/\alpha}$, with $s \in \mathbb{R}_+$, for some $\alpha \geqslant 1$.

A2. $g_{\mathcal{H}} \in T^r\left(L^2_{\rho_X}\right)$ with $r \geqslant 0$, and as a consequence $\|T^{-r}(g_{\mathcal{H}})\|_{L^2_{\rho_X}} < \infty$.

The two parameters r and α intuitively parametrize the strengths of our assumptions.

1) First assumption implies that the spectrum of the covariance operator (in decreasing order) decreases like $O(i^{-\alpha})$: a bigger α makes the assumption stronger.

2) In the second assumption, for a fixed α, a bigger r makes the assumption stronger, that is the function $g_{\mathcal{H}}$ is actually smoother. Indeed, for any $r < r'$, $T^{r'}\left(L^2_{\rho_X}\right) \subset T^r\left(L^2_{\rho_X}\right)$. Note that for $r = 1/2$, $T^{1/2}\left(L^2_{\rho_X}\right) = \mathcal{H}$; moreover, for $r \geqslant 1/2$, our best approximation function $g_{\mathcal{H}} \in \overline{\mathcal{H}}$ is in fact in $\mathcal{H}$, that is the optimization problem in the RKHS $\mathcal{H}$ is attained by a function of finite norm.

3 Stochastic approximation in Hilbert spaces

Algorithm. We consider the simple stochastic gradient descent algorithm, making one pass through the data, and averaging the result : we use the fact that for any independent pair of observations (x_n, y_n), the stochastic gradient $(y_n - \langle g, K_{x_n} \rangle_{\mathcal{H}}) K_{x_n}$ is an unbiased

estimate of $\nabla\varepsilon(g)$. Our algorithm minimizes directly generalization error, without extra regularization (early stopping is used to avoid over-fitting):

$$\left\{\begin{array}{lrcl} & g_0 & = & 0 \\ \forall n \geqslant 1 & g_n & = & g_{n-1} - \gamma_n \big[y_n - \langle g_{n-1}, K_{x_n}\rangle_{\mathcal{H}}\big] K_{x_n} = g_{n-1} - \gamma_n \big[y_n - g_{n-1}(x_n)\big] K_{x_n}. \end{array}\right.$$

$$\text{output} \quad \bar{g}_n = \frac{1}{n+1}\sum_{k=0}^{n} g_k.$$

where γ_n is the step-size. The recursion is computed using representants, we only compute the sequence of coefficients $(a_i)_{1\leqslant i\leqslant n}$ such that for any $n \geqslant 1$, $g_n = \sum_{i=1}^n a_i K_{x_i}$.

Running-time complexity. The running time complexity is $O(i)$ for iteration i and thus $O(n^2)$ after n steps. This is a serious limitation for practical applications. Some authors have considered expanding g_n on a subset of all (K_{x_i}), to bring down the complexity [9, 10, 11].

Learning rate. The sequence of step-sizes $(\gamma_i)_{1\leqslant i\leqslant n}$ may be :

1) either a subsequence of a universal sequence $(\gamma_i)_{i\in\mathbb{N}}$, we refer to this situation as the "online setting" and our bounds then hold for any of the iterates;

2) or a sequence of the type $\gamma_i = \Gamma(n)$ for $i \leqslant n$, which will be referred to as the "finite horizon setting": in this situation the number of samples is assumed to be known and fixed and we chose a constant step-size which may depend on this number. Our bound then hold only for the last iterate. Considering space limitation we will only give finite horizon bounds here, but the extension hold up to minor differences, see [1].

Extra regularity assumptions. We have to add an assumption to control the noise covariance : we assume that there exists $\sigma > 0$ such that $\mathbb{E}\left[\Xi \otimes \Xi\right] \preccurlyeq \sigma^2 T$, where $\preccurlyeq$ denotes the order between self-adjoint operators. This assumption is clearly satisfied in the well specified homoscedastic case.

3.1 Main results (finite horizon)

Our main theorem, in terms of generality, is the following :

Theorem 1. *Assume $\gamma_i = \gamma = \Gamma(n)$, for $1 \leqslant i \leqslant n$. If $\gamma R^2 \leqslant 1/4$:*

$$\mathbb{E}\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}} \leqslant \frac{4\sigma^2}{n}\left(1 + \mathrm{tr}(T^{\frac{1}{\alpha}})(\gamma n)^{\frac{1}{\alpha}}\right) + 4\frac{\|T^{-r} g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}}{\gamma^{2r} n^{2\min\{r,1\}}}.$$

We thus first get some guarantee on the consistency of our estimator, for any small enough constant step-size:

Corollary 1. *For any constant choice $\gamma_n = \gamma_0 < \frac{1}{2R^2}$, the prediction error of $\bar{g}_n$ converges to the one of $g_{\mathcal{H}}$, that is: $\mathbb{E}\left[\varepsilon\left(\bar{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \xrightarrow{n\to\infty} 0$.*

Proof. this bound is derived from a new error decomposition to control the different sources of error via algebraic calculations.

Bias/variance interpretation. The two main terms have a simple interpretation. The first one is a variance term, which shows the effect of the noise σ^2 on the error. It is bigger when σ gets bigger, and moreover it also gets bigger when γ is growing (bigger steps mean more variance). As for the second term, it is a bias term, which accounts for the distance of the initial choice (the null function in general) to the objective function. As a consequence, it is smaller when we make bigger steps.

Saturation. Dependence in n does not improve beyond $r > 1$ (while this assumption is stronger than $r = 1$): this phenomenon in known as saturation [12]. Improvements with $r > 1$ could be achieved by considering another type of averaging.

3.2 Statistical rate of prediction error in Hilbert spaces

We may now deduce the following corollary, with specific optimized values of γ:

Corollary 2 (Optimal constant γ)**.** *If $\frac{\alpha-1}{2\alpha} < r$ and $\Gamma(n) = \gamma_0\, n^{\frac{-2\alpha\min\{r,1\}-1+\alpha}{2\alpha\min\{r,1\}+1}}$, $\gamma_0 R^2 \leqslant 1/4$, we have:*

$$\mathbb{E}\left(\|\bar{g}_n - g_{\mathcal{H}}\|^2_{L^2_{\rho_X}}\right) \leqslant A\; n^{-\frac{2\alpha\min\{r,1\}}{2\alpha\min\{r,1\}+1}}.$$

with $A = 4\left(1 + \operatorname{tr}(T^{\frac{1}{\alpha}})\gamma_0^{\frac{1}{\alpha}}\right)\sigma^2 + \frac{4}{\gamma_0^{2r}}||T^{-r}g_{\mathcal{H}}||^2_{L^2_{\rho_X}}$.

A slightly different result holds for $\frac{\alpha-1}{2\alpha} > r$, see [1]. We can make the following observations:

Evolution with r and α. As it has been noticed above, a bigger α or r would be a stronger assumption. It is thus natural to get a rate which improves with a bigger α or r: the function $(\alpha, r) \mapsto \frac{2\alpha r}{2\alpha r+1}$ is increasing in both parameters.

Optimal rates. In some situations, our stochastic approximation framework leads to "optimal" rates of prediction from a statistical point of view. Indeed, in [4, Theorem 2] a corresponding minimax lower rate is given showing that no estimator can get a better rate of convergence for all objective functions than $n^{-\frac{2\alpha r}{2\alpha r+1}}$. Our algorithm thus enjoys this optimal rate. In Figure 1, we plot in the plan of coordinates α, δ (with $\delta = 2\alpha r+1$) our limit conditions concerning our assumptions. The region between the two green lines is the region for which the optimal rate of estimation is reached.

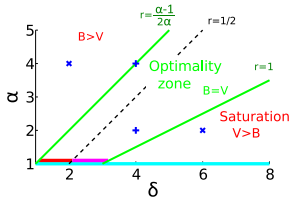

Figure 1

3.3 Adaptativity in Euclidean spaces

We can also derive the following corollary, which shows adaptativity in Euclidean spaces :

Corollary 3. *Assume $\mathcal{H}$ is a d-dimensional Euclidean space:*

$$\mathbb{E}\left[\varepsilon\left(\overline{g}_n\right) - \varepsilon(g_{\mathcal{H}})\right] \leqslant \min_{1\leqslant\alpha, \frac{-1}{2}\leqslant q\leqslant\frac{1}{2}} \left(16\frac{\sigma^2 \operatorname{tr}(T^{1/\alpha})(\gamma n)^{1/\alpha}}{n} + 8\frac{||T^{-q}g_{\mathcal{H}}||^2_{\mathcal{H}}}{(n\gamma)^{2q+1}}\right).$$

This results shows that SGD is adaptative to the regularity of the objective function and to the decay of the spectrum of the covariance matrix : it bridges the gap between different results from [3] and [13].

It shows for example that :

1) the variance term is always smaller than both $\frac{d}{n}(\alpha \to \infty)$ and $\gamma \operatorname{tr}(T)$ $(\alpha = 1)$;

2) similarly the bias is smaller than both $\frac{\|\theta\|^2_{\mathcal{H}}}{\gamma n}$ $(q = 0)$ and $\frac{\|T^{-1/2}\theta\|^2_{\mathcal{H}}}{(\gamma n)^2}$ $(q = 1/2)$.

This explains the robustness of SGD in high dimension, when $d >> n$, for example. It means that when the problem is easier, the algorithm will perform better.

Relationship to previous works. Similar algorithms have been studied before [5, 6, 14, 15, 16, 17], under various forms. A detailed discussion is given in [1].

4 Conclusion

In this paper, we have provided an analysis of averaged unregularized stochastic gradient methods for kernel-based least-squares regression. Our novel analysis allowed us to consider larger step-sizes, which in turn lead to optimal estimation rates for many settings of eigenvalue decay of the covariance operators and smoothness of the optimal prediction function. Moreover, it showed that this simple algorithm is indeed adaptative to the smoothness of the objective function, and to the decay of the spectrum of the covariance matrix.

Our work can be extended in a number of interesting ways, mainly: (a) while we obtain optimal convergence rates for a particular regime of kernels/objective functions, using different types of averaging (i.e., non uniform) may lead to optimal rates in other regimes. Also, (b) the running-time complexity of our stochastic approximation procedures is still quadratic in the number of samples n, which is unsatisfactory when n is large; by considering reduced set-methods [9, 10, 11], we could improve that dependency. Finally, (c) in order to obtain the optimal rates when the bias term dominates our generalization bounds, it would be interesting to combine our spectral analysis with recent accelerated versions of stochastic gradient descent which have been analyzed in the finite-dimensional setting [13].

References

[1] A. Dieuleveut and F. Bach. Non-parametric Stochastic Approximation with Large Step sizes. ArXiv e-prints, August 2014.

[2] H. Robbins and S. Monro. A stochastic approxiation method. The Annals of mathematical Statistics, 22(3):400–407, 1951.

[3] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). Advances in Neural Information Processing Systems (NIPS), 2013.

[4] A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. Foundations of Computational Mathematics, 7(3):331–368, 2007.

[5] L. Rosasco, A. Tacchetti, and S. Villa. Regularization by Early Stopping for Online Learning Algorithms. ArXiv e-prints, 2014.

[6] Y. Ying and M. Pontil. Online gradient descent learning algorithms. Foundations of Computational Mathematics, 5, 2008.

[7] S. Smale and F. Cucker. On the mathematical foundations of learning. Bulletin of the American Mathematical Society, 39(1):1–49, 2001.

[8] S. Smale and D-X. Zhou. Learning theory estimates via integral operators and their approximations. Constructive Approximation, 26(2):153–172, 2007.

[9] O. Dekel, S. Shalev-Shwartz, and Y. Singer. The Forgetron: A kernel-based perceptron on a fixed budget. In Adv. NIPS, 2005.

[10] A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. Journal of Machine Learning Research, 6:1579–1619, 2005.

[11] F. Bach. Sharp analysis of low-rank kernel matrix approximations. Proceedings of the International Conference on Learning Theory (COLT), 2012.

[12] H. W. Engl, M. Hanke, and Neubauer A. Regularization of inverse problems. Klüwer Academic Publishers, 1996.

[13] N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In Proceedings of the International Conference on Learning Theory (COLT), 2015.

[14] J. Kivinen, Smola A.J., and R. C. Williamson. Online learning with kernels. IEEE transactions on signal processing, 52(8):2165–2176, 2004.

[15] Y. Yao. A dynamic Theory of Learning. PhD thesis, University of California at Berkeley, 2006.

[16] P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths. ArXiv e-prints 1103.5538, 2011.

[17] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. ICML 2014 Proceedings of the twenty-first international conference on machine learning, 2004.