Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression

Aymeric Dieuleveut
aymeric.dieuleveut@ens.fr
INRIA - École Normale Supérieure, Paris

Nicolas Flammarion
nicolas.flammarion@ens.fr
INRIA - École Normale Supérieure, Paris

Francis Bach
francis.bach@ens.fr
INRIA - École Normale Supérieure, Paris

Abstract

We consider the optimization of a quadratic objective function whose gradients are only accessible through a stochastic oracle. We present the first algorithm that achieves jointly the optimal prediction error rates for least-squares regression, both in terms of forgetting of initial conditions and in terms of dependence on the noise and dimension of the problem, and prove dimensionless and tighter rates for a regularized version of this algorithm.

1 Introduction

Many supervised machine learning problems are naturally cast as the minimization of a smooth function defined on a Euclidean space. This includes least-squares regression, logistic regression (see, e.g., Hastie et al., 2009) or generalized linear models (McCullagh and Nelder, 1989). While small problems with few or low-dimensional input features may be solved precisely by many potential optimization algorithms (e.g., Newton method), large-scale problems with many high-dimensional features are typically solved with simple gradient-based iterative techniques whose per-iteration cost is small.

In this paper, we consider a quadratic objective function f whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. In this stochastic approximation framework (Robbins and Monro, 1951), it is known that two quantities dictate the behavior of various algorithms, namely the covariance matrix V of the noise in the gradients, and the deviation $\theta_0 - \theta_*$ between the initial point of the algorithm θ_0 and any of the global minimizer θ_* of f. This leads to a "bias/variance" decomposition (Bach and Moulines, 2013; Hsu et al., 2014) of the performance of most algorithms as the sum of two terms: (a) the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; while (b) the variance term characterizes the effect of the noise in the gradients, independently of the starting point, and with a term that is increasing in the covariance of the noise.

For quadratic functions with (a) a noise covariance matrix V which is proportional (with constant σ^2) to the Hessian of f (a situation which corresponds to least-squares regression) and (b) an initial point characterized by the norm $\|\theta_0 - \theta_*\|^2$, the optimal bias and variance terms are known *separately*. On the one hand, the optimal bias term after n iterations is proportional to $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$, where L is the largest eigenvalue of the Hessian of f. This rate is achieved by accelerated gradient descent (Nesterov, 1983, 2004), and is known to be optimal if the number of iterations n is less than the dimension d of the underlying predictors, but the algorithm is not robust to random or deterministic noise in the gradients (d'Aspremont, 2008; Schmidt et al., 2011; Devolder et al., 2014). On the other

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

hand, the optimal variance term is proportional to $\frac{\sigma^2 d}{n}$ (Tsybakov, 2003); it is known to be achieved by averaged gradient descent (Bach and Moulines, 2013), for which the bias term only achieves $\frac{L\|\theta_0-\theta_*\|^2}{n}$ instead of $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$.

Our first contribution in this paper is to analyze in Section 3 averaged *accelerated* gradient descent, showing that it attains optimal rates for *both the variance and the bias terms*. It shows that averaging is beneficial for accelerated techniques and provides a provable robustness to noise.

While optimal when measuring performance in terms of the dimension d and the initial distance to optimum $\|\theta_0 - \theta_*\|^2$, these rates are not adapted in many situations where either d is larger than the number of iterations n (i.e., the number of observations for regular stochastic gradient descent) or $L\|\theta_0 - \theta_*\|^2$ is much larger than n^2. Our second contribution is to provide in Section 4 an analysis of a new algorithm (based on some additional regularization) that can adapt our bounds to finer assumptions on $\theta_0 - \theta_*$ and the Hessian of the problem, leading in particular to dimension-free quantities that can thus be extended to the Hilbert space setting (in particular for non-parametric estimation).

2 Least-Squares Regression

Statistical Assumptions. We consider the following general setting: $\mathcal{H}$ is a d-dimensional Euclidean space with $d \geq 1$, the observations $(x_n, y_n) \in \mathcal{H} \times \mathbb{R}$, $n \geq 1$, are independent and identically distributed (i.i.d.), and such that $\mathbb{E}\|x_n\|^2$ and $\mathbb{E}y_n^2$ are finite. We consider the *least-squares regression* problem which is the minimization of the quadratic function $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$.

Covariance matrix: We denote by $\Sigma = \mathbb{E}(x_n \otimes x_n) \in \mathbb{R}^{d\times d}$ the population covariance matrix, which is the Hessian of f at all points. Without loss of generality, we can assume Σ invertible. This implies that all eigenvalues of Σ are strictly positive (but they may be arbitrarily small). We assume there exists $R > 0$ such that $\mathbb{E}\|x_n\|^2 x_n \otimes x_n \preccurlyeq R^2\Sigma$ where $A \preccurlyeq B$ means that $B - A$ is positive semi-definite. This assumption is satisfied, for example, for least-square regression with almost surely bounded data.

Eigenvalue decay: Most convergence bounds depend on the dimension d of $\mathcal{H}$. However it is possible to derive dimension-free and often tighter convergence rates by considering bounds depending on the value $\operatorname{tr}\Sigma^b$ for $b \in [0, 1]$. Given b, if we consider the eigenvalues of Σ ordered in decreasing order, which we denote by s_i, then $\operatorname{tr}\Sigma^b = \sum_i s_i^b$, and the eigenvalues decay For b going to 0 then $\operatorname{tr}\Sigma^b$ tends to d and we are back in the classical low-dimensional case. When $b = 1$, we simply get $\operatorname{tr}\Sigma = \mathbb{E}\|x_n\|^2$, which will correspond to the weakest assumption in our context.

Optimal predictor: The regression function $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$ always admits a global minimum $\theta_* = \Sigma^{-1}\mathbb{E}(y_n x_n)$. When initializing algorithms at $\theta_0 = 0$ or regularizing by the squared norm, rates of convergence generally depend on $\|\theta_*\|$, a quantity which could be arbitrarily large. However there exists a systematic upper-bound $\|\Sigma^{\frac{1}{2}}\theta_*\| \leq 2\sqrt{\mathbb{E}y_n^2}$. This leads naturally to the consideration of convergence bounds depending on $\|\Sigma^{r/2}\theta_*\|$ for $r \leq 1$.

Noise: We denote by $\varepsilon_n = y_n - \langle\theta_*, x_n\rangle$ the residual for which we have $\mathbb{E}[\varepsilon_n x_n] = 0$. Although we do not have $\mathbb{E}[\varepsilon_n|x_n] = 0$ in general unless the model is well-specified, we assume the noise to be a structured process such that there exists $\sigma > 0$ with $\mathbb{E}[\varepsilon_n^2 x_n \otimes x_n] \preccurlyeq \sigma^2\Sigma$. This assumption is satisfied for example for data almost surely bounded or when the model is well-specified.

Averaged Gradient Methods and Acceleration. We focus in this paper on stochastic gradient methods with acceleration for a quadratic function regularized by $\frac{\lambda}{2}\|\theta - \theta_0\|^2$. The regularization will be useful when deriving tighter convergence rates in Section 4, and it has the additional benefit of making the problem λ-strongly-convex.

Accelerated stochastic gradient descent is defined by an iterative system with two parameters (θ_n, ν_n) starting from $\theta_0 = \nu_0 \in \mathcal{H}$, and satisfying for $n \geq 1$,

$$\begin{aligned} \theta_n &= \nu_{n-1} - \gamma f_n'(\nu_{n-1}) - \gamma\lambda(\nu_{n-1} - \theta_0) \\ \nu_n &= \theta_n + \delta\big(\theta_n - \theta_{n-1}\big), \end{aligned} \tag{1}$$

with $\gamma, \delta \in \mathbb{R}^2$ and $f_n'(\theta_{n-1})$ an unbiased estimate on the gradient $f(\theta)$.

The *momentum* coefficient $\delta \in \mathbb{R}$ is chosen to accelerate the convergence rate (Nesterov, 1983; Beck and Teboulle, 2009) and has its roots in the heavy-ball algorithm from Polyak (1964). We

especially concentrate here, following Polyak and Juditsky (1992), on the average of the sequence $\bar{\theta}_n = \frac{1}{n+1} \sum_{i=0}^{n} \theta_n$,

Stochastic Oracles on the Gradient. Let $(\mathcal{F}_n)_{n \geq 0}$ be the increasing family of σ-fields that are generated by all variables (x_i, y_i) for $i \leq n$. The oracle we consider is the sum of the true gradient $f'(\theta)$ and an independent zero-mean noise that does not depend on θ[1]. Consequently it is of the form $f_n'(\theta) = f'(\theta) - \xi_n$ where the noise process ξ_n is $\mathcal{F}_n$-measurable with $\mathbb{E}[\xi_n | \mathcal{F}_{n-1}] = 0$ and $\mathbb{E}[\|\xi_n\|^2]$ is finite. Furthermore we also assume that there exists $\tau \in \mathbb{R}$ such that $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \tau^2 \Sigma$, that is, the noise has a particular structure adapted to least-squares regression.

3 Accelerated Stochastic Averaged Gradient Descent

We study the convergence of averaged *accelerated* stochastic gradient descent defined by Eq. (1) for $\lambda = 0$ and $\delta = 1$. It can be rewritten for the quadratic function f as a second-order iterative system with constant coefficients: $\theta_n = \big[I - \gamma \Sigma\big](2\theta_{n-1} - \theta_{n-2}) + \gamma y_n x_n$.

Theorem 1 *For any constant step-size γ, such that $\gamma \Sigma \preccurlyeq I$,*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq 36 \frac{\|\theta_0 - \theta_*\|^2}{\gamma (n+1)^2} + 8 \frac{\tau^2 d}{n+1}. \tag{2}$$

We can make the following observations:

- The first bound $\frac{1}{\gamma n^2} \|\theta_0 - \theta_*\|^2$ in Eq. (2) corresponds to the usual accelerated rate. It has been shown by Nesterov (2004) to be the optimal rate of convergence for optimizing a quadratic function with a first-order method that can access only to sequences of gradients when $n \leq d$. We recover by averaging an algorithm dedicated to strongly-convex function the traditional convergence rate for non-strongly convex functions.
- The second bound in Eq. (2) matches the optimal statistical performance $\frac{\tau^2 d}{n}$ over all estimators in $\mathcal{H}$ (Tsybakov, 2008) even without computational limits, in the sense that no estimator that uses the same information can improve upon this rate. Accordingly this algorithm achieves joint bias/variance optimality (when measured in terms of τ^2 and $\|\theta_0 - \theta_*\|^2$).
- We have the same rate of convergence for the bias when compared to the regular Nesterov acceleration without averaging studied by Flammarion and Bach (2015), which corresponds to choosing $\delta_n = 1 - 2/n$ for all n. However if the problem is μ-strongly convex, this latter was shown to also converge at the linear rate $O\big((1 - \gamma\mu)^n\big)$ and thus is adaptive to hidden strong-convexity (since the algorithm does not need to know μ to run), thus ends up converging faster than the rate $1/n^2$. This is confirmed in our experiments in Section 5.
- Overall, the bias term is improved whereas the variance term is not degraded and acceleration is thus robust to noise in the gradients. Thereby, while second-order iterative methods for optimizing quadratic functions in the singular case, such as conjugate gradient (Polyak, 1987, Section 6.1) are notoriously highly sensitive to noise, we are able to propose a version which is robust to stochastic noise.

4 Tighter Convergence Rates

We have seen in Corollary 1 above that the averaged accelerated gradient algorithm matches the lower bounds $\tau^2 d/n$ and $\frac{L}{n^2} \|\theta_0 - \theta_*\|^2$ for the prediction error. However the algorithm performs better in almost all cases except the worst-case scenarios corresponding to the lower bounds. For example the algorithm may still predict well when the dimension d is much bigger than n. Similarly the norm of the optimal predictor $\|\theta_*\|^2$ may be huge and the prediction still good, as gradient algorithms happen to be adaptive to the difficulty of the problem: indeed, if the problem is simpler, the convergence rate of the gradient algorithm will be improved. In this section, we provide such a theoretical guarantee.

We study the convergence of averaged *accelerated* stochastic gradient descent defined by Eq. (1) for $\lambda = (\gamma(n+1)^2)^{-1}$ and $\delta \in \big[1 - \frac{2}{n+2}, 1\big]$. We have the following theorem:

[1] this is different from the oracle usually considered in stochastic approximation (see Bach and Moulines (2013); Dieuleveut and Bach (2015)).

Theorem 2 *For any constant step-size γ, such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$,*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq \min_{r \in [0,1],\ b \in [0,1]} \left[74 \, \frac{\|\Sigma^{r/2}(\theta_0 - \theta_*)\|^2}{\gamma^{1-r}(n+1)^{2(1-r)}} + 8 \frac{\tau^2 \gamma^b \operatorname{tr}(\Sigma^b)}{(n+1)^{1-2b}} \right].$$

We can make the following observations:

- The algorithm is independent of r and b, thus all the bounds for different values of (r, b) are valid. This is a strong property of the algorithm, which is indeed adaptative to the regularity and the effective dimension of the problem (once γ is chosen). In situations in which either d is larger than n or $L\|\theta_0 - \theta_*\|^2$ is larger than n^2, the algorithm can still enjoy good convergence properties, by adapting to the best values of b and r.
- For $b = 0$ we recover the variance term of Corollary 1, but for $b > 0$ and fast decays of eigenvalues of Σ, the bound may be much smaller; note that we lose in the dependency in n, but typically, for large d, this can be advantageous.
- With r, b well chosen, we recover the optimal rate for non-parametric regression (Caponnetto and De Vito, 2007).

5 Experiments

We illustrate now our theoretical results on synthetic examples. For $d = 25$ we consider normally distributed inputs x_n with random covariance matrix Σ which has eigenvalues $1/i^3$, for $i = 1, \dots, d$, and random optimum θ_* and starting point θ_0 such that $\|\theta_0 - \theta_*\| = 1$. The outputs y_n are generated from a linear function with homoscedastic noise with unit signal to noise-ratio ($\sigma^2 = 1$), we take $R^2 = \operatorname{tr} \Sigma$ the average radius of the data and a step-size $\gamma = 1/R^2$ and $\lambda = 0$. The additive noise oracle is used. We show results averaged over 10 replications.

We compare the performance of averaged SGD (AvSGD), usual Nesterov acceleration for convex functions (AccSGD) and our novel averaged accelerated SGD (AvAccSGD)[2], on two different problems: one deterministic ($\|\theta_0 - \theta_*\| = 1$, $\sigma^2 = 0$) which will illustrate how the bias term behaves, and one purely stochastic ($\|\theta_0 - \theta_*\| = 0$, $\sigma^2 = 1$) which will illustrate how the variance term behaves. For the bias (left plot of Figure 1), AvSGD converges at speed $O(1/n)$, while AvAccSGD

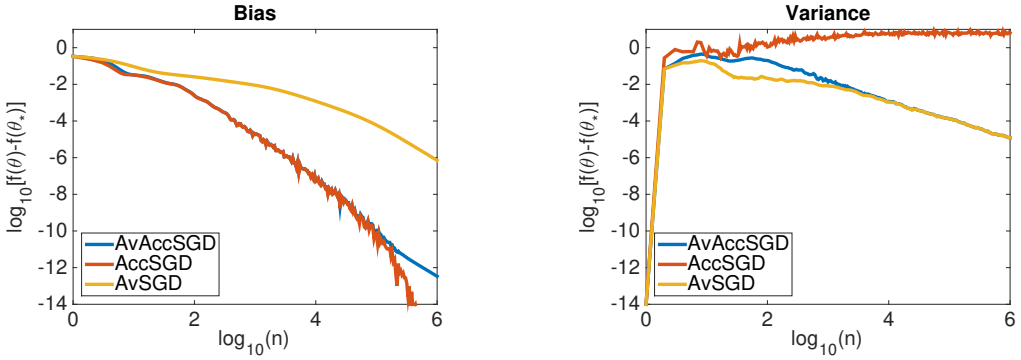

Figure 1: Synthetic problem ($d = 25$) and $\gamma = 1/R^2$. Left: Bias. Right: Variance.

and AccSGD converge both at speed $O(1/n^2)$. However, as mentioned in the observations following Theorem 1, AccSGD takes advantage of the hidden strong convexity of the quadratic function and starts converging linearly at the end. For the variance (right plot of Figure 1), AccSGD is not converging to the opltimum and keeps oscillating whereas AvSGD and AvAccSGD both converge to the optimum at a speed $O(1/n)$. However AvSGD remains slightly faster in the beginning.

Note that for small n, or when the bias $L\|\theta_0 - \theta_*\|^2/n^2$ is much bigger than the variance $\sigma^2 d/n$, the bias may have a stronger effect, although asymptotically, the variance always dominates. It is thus essential to have an algorithm which is optimal in both regimes; this is achieved by AvAccSGD.

[2] which is not the averaging of AccSGD because the momentum term is proportional to $1 - 3/n$ for AccSGD instead of being equal to 1 for AvAccSGD.

References

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

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

d'Aspremont, A. (2008). Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3):1171–1183.

Devolder, O., Glineur, F., and Nesterov, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2, Ser. A):37–75.

Dieuleveut, A. and Bach, F. (2015). Non-parametric stochastic approximation with large step sizes. *Annals of Statistics*.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition.

Hsu, D., Kakade, S. M., and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, second edition.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *{USSR} Computational Mathematics and Mathematical Physics*, 4(5):1–17.

Polyak, B. T. (1987). *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855.

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

Schmidt, M., Le Roux, N., and Bach, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In *Proceedings of the Annual Conference on Computational Learning Theory*.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer.