Learning with stochastic proximal gradient

Lorenzo Rosasco *
DIBRIS, Università di Genova
Via Dodecaneso, 35
16146 Genova, Italy
lrosasco@mit.edu

Silvia Villa, Băng Công Vũ
Laboratory for Computational and Statistical Learning
Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
{silvia.villa,cong.bang}@iit.it

Abstract

We consider composite function minimization problems where only a stochastic estimate of the gradient of the smooth term is available, and in particular regularized online learning. We derive novel finite sample bounds for the natural extension of the classical proximal gradient algorithm. Our approach allows to avoid averaging, a feature which is critical when considering sparsity based methods. Moreover, our results match those obtained by the stochastic extension of accelerated methods, hence suggesting that there is no advantage considering these variants in a stochastic setting.

1 Stochastic proximal gradient algorithm for composite optimization and learning

We consider the following general class of minimization problems

$$\min_{w\in\mathcal{H}} L(w) + R(w), \tag{1}$$

under the assumptions

- $L\colon \mathcal{H} \to \mathbb{R}$ is a differentiable convex function with β-Lipshcitz continuous gradient, i.e. for every v and w in $\mathcal{H}$, $\|\nabla L(v) - L(w)\| \leq \beta\|v - w\|$
- $R\colon \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous convex function (possibly nonsmooth)
- Problem 1 has at least one solution.

This class of optimization problems arises naturally in regularization schemes where one component is a data fitting term and the other a regularizer, see for example [6, 16]. To solve Problem 1, first order methods have recently been widely applied. In particular, proximal gradient algorithms (a.k.a. forward-backward splitting algorithms) and their accelerated variants have received considerable attention (see [8, 19, 3] and references therein). These algorithms are easy to implement and suitable for solving high dimensional problems thanks to the low memory requirement of each iteration. Interestingly, proximal splitting algorithms separate the contribution of each component at every

*Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology, Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

iteration: the proximity operator defined by the non smooth term is applied to a gradient descent step for the smooth term. Therefore, each iteration requires the computation of the proximity operator of R, that is

$$\mathrm{prox}_R \colon \mathcal{H} \to \mathcal{H}, \quad \mathrm{prox}_R(w) = \operatorname*{argmin}_{v \in \mathcal{H}} R(v) + \frac{1}{2}\|w - v\|^2. \tag{2}$$

Throughout this paper, we assume implicitly that the closed-form expression of the proximity operator of R is available, or that it can be cheaply computed. On the other hand, we suppose that the gradient is intractable, so that in the algorithm the gradient of L is replaced by a stochastic approximation. This latter situation is particularly relevant in statistical learning, where we have to minimize an expected objective function from random samples. In this context, iterative algorithms, where only one gradient estimate is used in each step, are often referred to as online learning algorithms. More generally, the situation where only stochastic gradient estimates are available is important in stochastic optimization, where iterative algorithms can be seen as a form of stochastic approximation.

More precisely, we study the following stochastic proximal gradient (SPG) algorithm.

SPG Algorithm. *Let $(\gamma_n)_{n\in\mathbb{N}^*}$ be a strictly positive sequence, let $(\lambda_n)_{n\in\mathbb{N}^*}$ be a sequence in $[0,1]$, and let $(\mathrm{G}_n)_{n\in\mathbb{N}^*}$ be a $\mathcal{H}$-valued random process such that $(\forall n \in \mathbb{N}^*)\ \mathbb{E}[\|\mathrm{G}_n\|^2] < +\infty$. Fix w_1 a $\mathcal{H}$-valued integrable vector with $\mathbb{E}[\|w_1\|^2] < +\infty$ and set*

$$(\forall n \in \mathbb{N}^*) \quad \left\lfloor \begin{array}{l} u_n = \mathrm{prox}_{\gamma_n R}(w_n - \gamma_n \mathrm{G}_n) \\ w_{n+1} = (1-\lambda_n) w_n + \lambda_n u_n. \end{array} \right. \tag{3}$$

The following conditions will be considered for the filtration $(\mathcal{A}_n)_{n\in\mathbb{N}^*}$ with $\mathcal{A}_n = \sigma(w_1, \ldots, w_n)$.

(A1) For every $n \in \mathbb{N}^*$, $\mathbb{E}\left[\mathrm{G}_n | \mathcal{A}_n\right] = \nabla L(w_n)$.

(A2) For every $n \in \mathbb{N}^*$, there exists $\sigma \in]0, +\infty[$ such that

$$\mathbb{E}\left[\|\mathrm{G}_n - \nabla L(w_n)\|^2 | \mathcal{A}_n\right] \leq \sigma^2 (1 + \|\nabla L(w_n)\|^2) \tag{4}$$

(A3) There exists $\epsilon \in]0, +\infty[$ such that $(\forall n \in \mathbb{N}^*)\ 0 < \gamma_n \leq \dfrac{1-\epsilon}{\beta(1+2\sigma^2)}$

(A4) For any solution $\overline{w}$ of Problem 1 assume that

$$\sum_{n\in\mathbb{N}^*} \lambda_n \gamma_n = +\infty \quad \text{and} \quad \sum_{n\in\mathbb{N}^*} \lambda_n \gamma_n^2 < +\infty. \tag{5}$$

Condition (A1) means that, at each iteration n, G_n is an unbiased estimate of the gradient of the smooth term. Condition (A2) is weaker than typical conditions used in the analysis of stochastic (sub)gradient algorithms, namely boundedness of the sequence $(\mathbb{E}[\|\mathrm{G}_n\|^2|\mathcal{A}_n])_{n\in\mathbb{N}^*}$ (see [17]) or even boundedness of $(\|\mathrm{G}_n\|^2)_{n\in\mathbb{N}^*}$ (see [10]). We note that this last requirement on the entire space is not compatible with the assumption of strong convexity, because the gradient is necessarily not uniformly bounded, therefore the use of the more general condition (A2) is needed in this case. Conditions such as (A3) and (A4) are, respectively, widely used in the deterministic setting and in stochastic optimization. Assumption (A3) is more restrictive that the one usually assumed in the deterministic setting, that is $(\forall n \in \mathbb{N}^*)\ \gamma_n \leq (2-\epsilon)/\beta$. We also note that when $(\lambda_n)_{n\in\mathbb{N}^*}$ is bounded away from zero, (A4) implies (A3) for n large enough. Finally, in our case, the step-size is required to converge to zero, while it is typically bounded away from zero in the study of deterministic proximal forward-backward splitting algorithm [8]. Viceversa, we allow for nonvanishing errors, while in the deterministic setting summability of $(\|G_n - \nabla L(w_n)\|)_{n\in\mathbb{N}^*}$ is required.

We next describe two settings particularly relevant in machine learning, in which SPG algorithm can be applied. Consider two measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ and assume there is a probability measure ρ on $\mathcal{X} \times \mathcal{Y}$. The measure ρ is fixed but known only through a training set $\mathbf{z} = (x_n, y_n)_{1\leq n\leq m} \in (\mathcal{X}\times\mathcal{Y})^m$ of samples i.i.d with respect to ρ. Consider a loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to [0, +\infty[$ and a hypothesis space $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$, e.g. a reproducing kernel Hilbert space. Assume that, for every $(x, y) \in \mathcal{X}\times\mathcal{Y}$, $\ell(y, \cdot)$ is a convex differentiable function with Lipschitz continuous gradient, examples being the squared or the logistic loss. Let R be convex, proper, and lower semicontinuous.

Example 1 (**Minimization of the (regularized) Risk).** A key problem in this context is

$$\underset{w\in\mathcal{H}}{\text{minimize}} \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, w(x))\, d\rho + R(w). \tag{6}$$

For every $x \in \mathcal{X}$, let $\mathrm{ev}_x \colon \mathcal{H} \to \mathcal{Y}$ defined as $\mathrm{ev}_x(w) = w(x)$ be the evaluation functional. By setting $(\forall n \in \mathbb{N}^*)$ $\mathrm{G}_n = \mathrm{ev}_{x_n}^* \nabla\ell(y_n, \cdot)(w_n(x_n))$ and $\mathcal{A}_n = \sigma((x_1, y_1), \ldots, (x_n, y_n))$, then (A1) holds. If assumption (A2) is satisfied, SPG algorithm can be applied for suitable choices of the step-size and the relaxation parameters $(\lambda_n)_{1\leq n\leq m\}}$. Note that only m steps of the algorithm can be taken.

Example 2 (**Minimization of the empirical risk).** The minimization of the regularized empirical risk

$$\underset{w\in\mathcal{H}}{\text{minimize}} \frac{1}{m}\sum_{i=1}^{m} \ell(y_i, w(x_i)) + R(w), \tag{7}$$

is often a key step to build a learning algorithm. This problem is a special case of Problem 1 with $L(w) = \sum_{i=1}^{m} \ell(y_i, w(x_i))$, and is especially of interest when m is very large and we know the exact gradient of each component in the sum. For every $n \in \mathbb{N}^*$ the stochastic estimate of the gradient of L is then defined as

$$(\forall n \in \mathbb{N}) \quad \mathrm{G}_n = \mathrm{ev}_{x_{\mathrm{i}(n)}}^* \nabla\ell(y_{\mathrm{i}(n)}, \cdot)(w_n(x_{\mathrm{i}(n)})), \tag{8}$$

where $(\mathrm{i}(n))_{n\in\mathbb{N}^*}$ is a random process of independent random variables uniformly distributed on $\{1, \ldots, m\}$, see [4, 5]. Clearly $(A1)$ holds. If $(A2)$ holds, SPG can be applied.

For loss functions which are twice differentiable in their second argument, it easy to see that the the maximum eigenvalue of the Hessian is a Lipschitz constant of the gradient. The term R can be seen as a regularizer/penalty encoding some prior information about the learning problem. Examples of convex, non-differentiable penalties include sparsity inducing penalties such as the ℓ_1 norm, as well as more complex structured sparsity penalties [16]. Stronger convexity properties can be obtained considering an elastic net penalty [9], that is adding a small strongly convex term to the sparsity inducing penalty. Clearly, the latter term would not be necessary if the risk in Problem 6 (or the empirical risk in (7)) is strongly convex. However, this latter requirement depends on the probability measure ρ and is typically not satisfied when considering high (possibly infinite) dimensional settings.

2 Theoretical and empirical analysis

In this section, we denote by $\overline{w}$ a solution of Problem 1 and provide a of convergence for $\mathbb{E}[\|w_n - \overline{w}\|^2]$. This result follows from nonasymptotic bounds to the nonsmooth case the bound obtained in [2, Theorem 1] for stochastic gradient descent. The following assumption is considered throughout this section.

Assumption 1. *The function L is μ-strongly convex and R is ν-strongly convex, for some $\mu \in [0, +\infty[$ and $\nu \in [0, +\infty[$, with $\mu + \nu > 0$.*

Note that, we do not assume both L and R to be strongly convex, indeed the constants μ and ν can be zero, but require that only one of the two is. This implies that Problem 1 has a unique solution, say $\overline{w}$.

Theorem 1. *Assume that conditions $(A1), (A2), (A3)$ and Assumption 1 are satisfied. Suppose that there exists $\underline{\lambda} \in]0, +\infty[$ such that $\inf_{n\in\mathbb{N}^*} \lambda_n \geq \underline{\lambda}$. Let $c_1 \in]0, +\infty[$ and let $\theta \in]0, 1]$. Suppose that, for every $n \in \mathbb{N}$, $\gamma_n = c_1 n^{-\theta}$. Set $c = \big(2c_1\underline{\lambda}(\nu + \mu\varepsilon)\big)/(1+\nu)^2$ and let n_0 be the smallest integer such that $n_0 > 1$, and $\max\{c, c_1\}{n_0}^{-\theta} \leq 1$. Then, for every $n \geq 2n_0$,*

$$\mathbb{E}[\|w_n - \overline{w}\|^2] = \begin{cases} O(n^{-\theta}) & \text{if } \theta \in]0, 1[, \\ O(n^{-c}) + O(n^{-1}) & \text{if } \theta = 1. \end{cases} \tag{9}$$

Thus, if $\theta = 1$ and c_1 is chosen such that $c > 1$, then $\mathbb{E}[\|w_n - \overline{w}\|^2] = O(n^{-1})$. More precisely, if $\theta = 1$, $\lambda_n = 1 = \underline{\lambda}$ for every $n \in \mathbb{N}^*$, and $c_1 = (1+\nu)^2/\big(\underline{\lambda}(\nu + \mu\varepsilon)\big) > 2$, then $c = 2$, $n_0 = \max\{2, c_1\}$, and

$$\mathbb{E}[\|w_n - \overline{w}\|^2] \leq \frac{n_0^2\mathbb{E}[\|w_{n_0} - \overline{w}\|^2]}{(n+1)^2} + \frac{8\sigma^2(1 + \|\nabla L(\overline{w})\|)(1+\nu)^4}{\underline{\lambda}^2\,(\mu\epsilon + \nu)^2} \tag{10}$$

We note that a recent technical report [1] also analyzes a stochastic proximal gradient method (without the relaxation step) and its accelerated variant. Almost sure convergence of the iterates (without averaging) is proved under uniqueness of the minimizer, but under assumptions different from ours: continuity of the objective function– thus excluding constrained smooth optimization– and boundedness of the iterates. Convergence rates for the iterates without averaging are derived, but only for the accelerated method. Finally, we note that convergence of the iterates of stochastic proximal gradient has been recently obtained from the analysis of convergence of stochastic fixed point algorithms presented in the recent preprint [7]. However, this latter results is derived from summability assumptions on the errors of the stochastic estimates which are usually not satisfied in the machine learning setting. The FOBOS algorithm in [10] is the closest approach to the one we consider, the main two differences being 1) we consider an additional relaxation step which may lead to accelerations, and especially 2) we do not consider averaging of the iterates. This latter point is important, since averaging can have a detrimental effect. Indeed, non-smooth problems often arise in applications where sparsity of the solution is of interest, and it is easy to see that averaging prevent the solution to be sparse [15, 22]. Moreover, as noted in [20] and [21], averaging can have a negative impact on the convergence rate in the strongly convex case. Indeed, in this paper we improve the error bound in [10] in this latter case. The best rate of convergence $O(1/n)$ is obtained for $\gamma_n = c_1/n$. There are other stochastic first order methods achieving the same rate of convergence for the iterates in the strongly convex case, see e.g. [1, 12, 11, 13, 22, 15]. Indeed, the rate we obtain is the rate that can be obtained by the optimal (in the sense of [18]) convergence rate on the function values. Among the mentioned methods those in [1, 11, 15] belong to the class of accelerated proximal gradient methods. Our result shows that, in the strongly convex case, the rate of convergence of the iterates is the same in the accelerated and non accelerated case. In addition, if sparsity is the main interest, we highlight that many of the algorithms discussed above (e.g. [1, 11, 12, 22]) involve some form of averaging or linear combination which prevent sparsity of the iterates, as it is discussed in [15]. Our result shows that in this case averaging is not needed, since the iterates themselves are convergent.

Next, we compare also numercally the proposed method with other state-of-the-art stochastic first order methods: an accelerated stochastic proximal gradient method, called SAGE [14] and the FOBOS algorithm [10]. We consider a regression problem with random design: for a suitably chosen finite dictionary of real valued functions $(\phi_k)_{1\leq k\leq p}$ defined on an interval, the labels are computed using a noise-corrupted regression function, namely

$$(\forall i \in \{1,\dots,N\}) \quad y_i = \sum_{k=1}^{p} \overline{w}_k \phi_k(x_i) + \epsilon_i, \tag{11}$$

where $(\overline{w_k})_{1\leq k\leq p} \in \mathbb{R}^p$ and ϵ_i is an additive noise $\epsilon_i \sim \mathcal{N}(0, 0.3)$. We considered two dictionaries (polynomial and trigonometric). On the regression problem with the polynomial dictionary, SAGE is performing the best, while on the trigonometric dictionary, SPG is the fastest. FOBOS shows a more regular behavior but slower convergence rate. Convergence of the iterations is displayed in Figure 1. We addressed also the problem of sparsity. Starting from an original signal with 1024

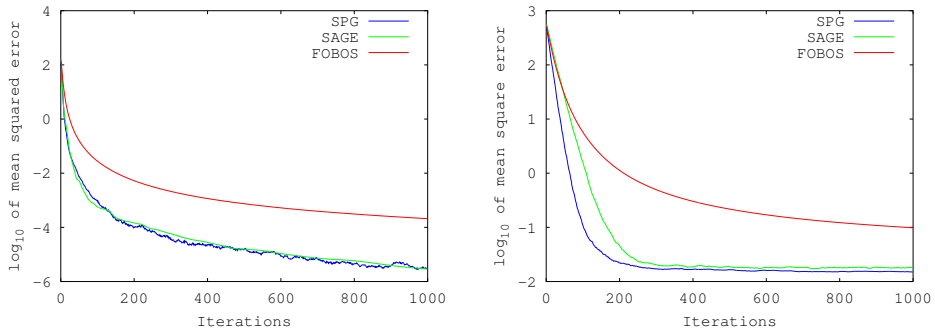

Figure 1: The convergence of the iterations to the optimal solution with polynomial dictionary (left) and with trigonometric dictionary (right).

components having 993 zero components, after few iteratons both SAGE ans SPG generate sparse

iterations (937 zero components), while the averaging procedure in FOBOS generated a vector with an increasing number of nonzero components, which was 438 after the last iteration.

References

[1] Y. F. Atchade, G. Fort, and E. Moulines. On stochastic proximal gradient algorithms. *arXiv:1402.2365*, February 2014.

[2] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Proceedings NIPS*, 2011.

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

[4] D. P. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.

[5] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, page 85, 2011.

[6] P. L. Combettes and J.-C. Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases. *SIAM J. Optim.*, 18(4):1351–1376, 2007.

[7] P. L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, 2014.

[8] P. L. Combettes and Valérie R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200 (electronic), 2005.

[9] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *J. Complexity*, 25:201–230, 2009.

[10] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.

[11] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[12] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512, 2014.

[13] A. Juditski and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

[14] J. T. Kwok, C. Hu, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 781–789, 2009.

[15] Q. Lin, X. Chen, and J. Peña. A sparsity preserving stochastic gradient methods for sparse regression. *Computational Optimization and Applications*, to appear, 2014.

[16] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa. Solving structured sparsity regularization with proximal methods. In *Machine Learning and Knowledge discovery in Databases European Conference, ECML PKDD 2010*, pages 418–433, Barcelona, Spain, 2010. Springer.

[17] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.

[18] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Intersci. Ser. Discrete Math. 15. John Wiley, New York, 1983.

[19] Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Paper 2007/76, Catholic University of Louvain, September 2007.

[20] A. Rakhlin, O. Shamir, and K. Sridaran. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[21] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[22] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.