Linearly Convergent Stochastic Heavy Ball Method for Minimizing Generalization Error

Nicolas Loizou n.loizou@sms.ed.ac.uk
University of Edinburgh, United Kingdom
Peter Richtárik peter.richtarik@ed.ac.uk
KAUST, Kingdom of Saudi Arabia
University of Edinburgh, United Kingdom

Abstract

In this work we establish the first linear convergence result for the stochastic heavy ball method. The method performs SGD steps with a fixed stepsize, amended by a heavy ball momentum term. In the analysis, we focus on minimizing the expected loss and not on finite-sum minimization, which is typically a much harder problem. While in the analysis we constrain ourselves to quadratic loss, the overall objective is not necessarily strongly convex.

1 Introduction

In this paper we study the stochastic optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\mathbf{S}\sim\mathcal{D}}[f_{\mathbf{S}}(x)] \tag{1}$$

where $\mathbf{A}\in\mathbb{R}^{m\times d}$ is a data matrix, $b\in\mathbb{R}^m$ is a vector of labels, $\mathbf{S}$ is a matrix with m rows (and arbitrary number of columns, e.g., 1), $\mathcal{D}$ is a distribution over such matrices and $f_{\mathbf{S}}(x) := \frac{1}{2}\|\mathbf{A}x-b\|^2_{\mathbf{H}}$ is a least-squares function with respect to a random pseudo-norm defined by a specific symmetric positive semidefinite matrix $\mathbf{H}$ which depends on $\mathbf{A}$ and the random matrix $\mathbf{S}$. In particular, $\|y\|^2_{\mathbf{H}} := y^\top\mathbf{H}y$ and $\mathbf{H} := \mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{A}^\top\mathbf{S})^\dagger\mathbf{S}^\top$, where $\dagger$ denotes the Moore-Penrose pseudoinverse. Note that the function f is finite if and only if $\mathbb{E}_{\mathbf{S}\sim\mathcal{D}}[\mathbf{H}]$ exists and is finite. Hence, we assume this throughout this paper.

Problem (1) was first proposed in [11], where the authors focus on stochastic reformulations of a consistent linear system $\mathbf{A}x = b$. The authors further give necessary and sufficient conditions on $\mathcal{D}$ for the set of solutions of (1) to be equal to the set of solutions of the linear system $\mathbf{A}x = b$; a property for which the term *exactness* was coined in [11]. Exactness conditions are very weak, allowing $\mathcal{D}$ to be virtually any distribution of random matrices. For instance, a sufficient condition for exactness is for the matrix $\mathbb{E}[\mathbf{H}]$ to be positive definite. This is indeed a weak condition since it is easy to see that this matrix is symmetric and positive semidefinite without the need to invoke any assumptions; simply by design. We refer the reader to [11] for more insights into the reformulation (1), its properties and other equivalent reformulations (e.g., stochastic fixed point problem, probabilistic intersection problem, and stochastic linear system).

In [11], the authors consider solving (1) via stochastic gradient descent (SGD)

$$x_{k+1} = x_k - \omega\nabla f_{\mathbf{S}_k}(x_k), \tag{2}$$

where $\omega > 0$ is a fixed stepsize and $\mathbf{S}_k$ is sampled afresh in each iteration from $\mathcal{D}$. It is shown that, SGD converges to an x_* which satisfies

$$x_* = \text{argmin}_{x\in\mathbb{R}^d}\tfrac{1}{2}\|x-x_0\|^2 \quad \text{subject to} \quad \mathbf{A}x = b, \tag{3}$$

where x_0 is the starting point. It was observed that, surprisingly, SGD is in this setting equivalent to the stochastic (pseudo)-Newton method, and the stochastic proximal point method, and that it converges at a linear rate despite the following obstacles: f is not necessarily strongly convex, (1) is not a finite-sum problem, and a fixed stepsize ω is used.

1.1 Contributions

In this paper we take an alternative route, and develop a *stochastic* variant of the *heavy ball method* for solving the stochastic optimization problem (1). Applied to (1), the classical heavy ball method of Polyak [9, 10], with

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

constant stepsize $\omega > 0$ and constant momentum parameter $\beta \geq 0$, takes the form

$$x_{k+1} = x_k - \omega \nabla f(x_k) + \beta(x_k - x_{k-1}) \tag{4}$$

This method introduces the momentum term $\beta(x_k - x_{k-1})$ into the gradient descent method to achieve acceleration.

Our stochastic variant of the heavy ball method, which we henceforth simply refer to by the name *stochastic heavy ball method (SHB)*, replaces the (costly) computation of the gradient by an unbiased estimator of the gradient ("stochastic gradient") which is hopefully much cheaper to compute:

$$\boxed{x_{k+1} = x_k - \omega \nabla f_{\mathbf{S}_k}(x_k) + \beta(x_k - x_{k-1})} \tag{5}$$

We establish global linear convergence (in expectation) of the iterates and function values: $\mathbb{E}[\|x_k - x_*\|^2] \to 0$ (L2 convergence) and $\mathbb{E}[f(x_k)] \to 0$. Without the exactness assumption we prove that $\mathbb{E}[f(\hat{x}_k)] = \mathcal{O}(1/k)$, where $\hat{x}_k = \frac{1}{k}\sum_{t=0}^{k-1} x_t$ is the Cesaro average. Finally, we study the convergence of the expected iterates (L1 convergence), $\|\mathbb{E}[x_k - x_*]\|^2 \to 0$, and establish global accelerated linear rate. That is, this quantity falls below ϵ after $\mathcal{O}\left((\lambda_{\max}/\lambda_{\min}^+)^{1/2}\log(1/\epsilon)\right)$ iterations, where $\lambda_{\max}$ (resp. $\lambda_{\min}^+$) are the largest (resp. smallest nonzero) eigenvalues of: $\nabla^2 f(x) = \mathbf{A}^\top \mathbb{E}_{\mathbf{S}\sim\mathcal{D}}[\mathbf{H}]\mathbf{A}$. It turns out that all eigenvalues of $\nabla^2 f(x)$ belong to the interval $[0, 1]$.

1.2 Related Work

Stochastic variants of heavy ball method have been employed widely in practice, especially in the area of deep learning [14, 15, 8]. Despite the popularity of the method both in convex and non-convex optimization its convergence properties are not very well understood. Recent papers that provide complexity analysis of SHB (in different setting than ours) include [16] and [3]. In [16] the authors analyzed SHB for general Lipshitz continuous convex objective functions (with bounded variance) and proved the *sublinear* rate $O(1/\sqrt{k})$. In [3], a complexity analysis is provided for the case of quadratic strongly convex smooth coercive functions. A *sublinear* convergence rate $O(1/k^\beta)$, where $\beta \in (0, 1)$, was proved. In contrast to our results, where we assume fixed stepsize ω, both papers analyze SHB with diminishing stepsizes. For our problem, variance reduction methods like SVRG [5], S2GD [7], mS2GD [6], SAG [12] and SAGA [2] are not necessary. To the best of our knowledge, *our work provides the first linear convergence rate for SHB in any setting.*

2 Convergence Results

In this section we state our convergence results for SHB.

2.1 $L2$ convergence: linear rate

We study L2 convergence of SHB; that is, we study the convergence of the quantity $\mathbb{E}[\|x_k - x_*\|^2]$ to zero. We show that for a range of stepsize parameters $\omega > 0$ and momentum parameters $\beta \geq 0$, SHB enjoys *global non-asymptotic linear convergence rate*. As a corollary of L2 convergence, we obtain convergence of the expected function values.

Theorem 1. *Choose $x_0 = x_1 \in \mathbb{R}^d$. Assume exactness. Let $\{x_k\}_{k=0}^\infty$ be the sequence of random iterates produced by SHB. Assume $0 < \omega < 2$ and $\beta \geq 0$ and that the expressions*

$$a_1 := 1 + 3\beta + 2\beta^2 - (\omega(2-\omega) + \omega\beta)\lambda_{\min}^+, \qquad \text{and} \qquad a_2 := \beta + 2\beta^2 + \omega\beta\lambda_{\max}$$

satisfy $a_1 + a_2 < 1$. Let x_* be the projection of x_0 onto $\{x \ : \ \mathbf{A}x = b\}$. Then

$$\mathbb{E}[\|x_k - x_*\|^2] \leq q^k(1+\delta)\|x_0 - x_*\|^2 \tag{6}$$

and

$$\mathbb{E}[f(x_k)] \leq q^k \tfrac{\lambda_{\max}}{2}(1+\delta)\|x_0 - x_*\|^2,$$

where $q = \frac{a_1 + \sqrt{a_1^2 + 4a_2}}{2}$ and $\delta = q - a_1$. Moreover, $a_1 + a_2 \leq q < 1$.

Remark 1. *In the above theorem we obtain global linear rate. To the best of our knowledge, this is the first time that linear rate is established for a stochastic variant of the heavy ball method in any setting. All existing results are sublinear.*

Remark 2. *If we choose $\omega \in (0,2)$, then the condition $a_1 + a_2 < 1$ is satisfied for all*

$$0 \leq \beta < \tfrac{1}{8}\left(-4 + \omega\lambda_{\min}^{+} - \omega\lambda_{\max} + \sqrt{(4 - \omega\lambda_{\min}^{+} + \omega\lambda_{\max})^2 + 16\omega(2-\omega)\lambda_{\min}^{+}}\right).$$

Remark 3. *If $\beta = 0$, SHB reduces to the "basic method" in [11] (SGD with constant stepsize). In this special case, $q = 1 - \omega(2-\omega)\lambda_{\min}^{+}$, which is the rate established in [11]. Hence, our result is more general.*

Remark 4. *Let $q(\beta)$ be the rate as a function of β. Note that since $\beta \geq 0$, we have*

$$q(\beta) \geq a_1 + a_2 = 1 + 4\beta + 4\beta^2 + \omega\beta(\lambda_{\max} - \lambda_{\min}^{+}) - \omega(2-\omega)\lambda_{\min}^{+} \geq 1 - \omega(2-\omega)\lambda_{\min}^{+} = q(0). \quad (7)$$

Clearly, the lower bound on q is an increasing function of β. Also, for any β the rate is always inferior to that of SGD ($\beta = 0$). It is an open problem whether one can prove a strictly better rate for SHB than for SGD.

2.2 Cesaro average: sublinear rate without exactness assumption

In this section we present convergence results for function values computed at the Cesaro average of all past iterates. Again, our results are global in nature. To the best of our knowledge, an analysis of the Cesaro average for the SHB with $O(1/k)$ rate was not established before for any class of functions. Moreover, the result holds without the exactness assumption.

Theorem 2. *Choose $x_0 = x_1$ and let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by SHB, where the momentum parameter $0 \leq \beta < 1$ and relaxation parameter (stepsize) $\omega > 0$ satisfy $\omega + 2\beta < 2$. Let x_* be any vector satisfying $f(x_*) = 0$. If we let $\hat{x}_k = \frac{1}{k}\sum_{t=1}^{k} x_t$, then*

$$\mathbb{E}[f(\hat{x}_k)] \leq \frac{(1-\beta)^2\|x_0 - x_*\|^2 + 2\omega\beta f(x_0)}{2\omega(2 - 2\beta - \omega)k}.$$

Remark 5. *In the special case of $\beta = 0$ we have $\mathbb{E}[f(\hat{x}_k)] \leq \frac{\|x_0 - x_*\|^2}{2\omega(2-\omega)k}$, which is the convergence rate for Cesaro average of the "basic method" analyzed in [11].*

2.3 $L1$ convergence: accelerated linear rate

In this section we show that by a proper combination of the stepsize parameter ω and the momentum parameter β the proposed algorithm enjoys *accelerated linear convergence* rate with respect to the expected iterates.

Theorem 3. *Assume exactness. Let $\{x_k\}_{k=0}^{\infty}$ be the sequence of random iterates produced SHB, started with $x_0, x_1 \in \mathbb{R}^d$ satisfying the relation $x_0 - x_1 \in \mathrm{Range}(\mathbf{A}^{\top})$, with stepsize parameter $0 < \omega \leq 1/\lambda_{\max}$ and momentum parameter $\left(1 - (\omega\lambda_{\min}^{+})^{1/2}\right)^2 < \beta < 1$. Then there exists constant $C > 0$ such that for all $k \geq 0$ we have $\|\mathbb{E}[x_k - x_*]\|^2 \leq \beta^k C$.*

(i) If we choose $\omega = 1$ and $\beta = \left(1 - \sqrt{0.99\lambda_{\min}^{+}}\right)^2$, then

$$\|\mathbb{E}[x_k - x_*]\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{0.99\lambda_{\min}^{+}}\right)^{2k} C$$

and the iteration complexity becomes $\mathcal{O}\left(\sqrt{1/\lambda_{\min}^{+}}\log(1/\epsilon)\right)$.

(ii) If we choose $\omega = 1/\lambda_{\max}$ and $\beta = \left(1 - \sqrt{0.99\lambda_{\min}^{+}/\lambda_{\max}}\right)^2$, then

$$\|\mathbb{E}[x_k - x_*]\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{0.99\lambda_{\min}^{+}/\lambda_{\max}}\right)^{2k} C$$

and the iteration complexity becomes $\mathcal{O}\left(\sqrt{\lambda_{\max}/\lambda_{\min}^{+}}\log(1/\epsilon)\right)$

Note that the convergence factor is precisely equal to the value of the momentum parameter.

Remark 6. *Let x be any random vector in $\mathbb{R}^d$ with finite mean $\mathbb{E}[x]$, and $x_* \in \mathbb{R}^d$ be any reference vector (for instance, any solution of $\mathbf{A}x = b$). Then we have the identity (see, for instance [4])*

$$\mathbb{E}[\|x - x_*\|^2] = \|\mathbb{E}[x - x_*]\|^2 + \mathbb{E}[\|x - \mathbb{E}[x]\|^2].$$

This means that the quantity $\mathbb{E}[\|x - x_*\|^2]$ appearing in our L2 convergence result (Theorem 1) is larger than $\|\mathbb{E}[x - x_*]\|^2$ appearing in the L1 convergence result (Theorem 3), and hence harder to push to zero. As a corollary, L2 convergence implies L1 convergence. However, note that in Theorem 3 we have established an accelerated *rate.*

3 Experiments

In this section we present a preliminary experiment to evaluate the performance of the SHB for solving the stochastic optimization problem (1). Matrices $\mathbf{A}$ are picked from the LIBSVM library [1]. To ensure consistency of the linear system, we take the optimal solution $x_* \in \mathbb{R}^d$ to be i.i.d $\mathcal{N}(0, 1)$ and the right hand sight is set to $b = \mathbf{A}x_*$. In each iteration, the random matrix is chosen as $\mathbf{S} = e_i \in \mathbb{R}^n$ with probability $p_i = \|\mathbf{A}_{i:}\|^2/\|\mathbf{A}\|_F^2$. Here e_i is the unit coordinate vector in $\mathbb{R}^n$. In this setup the update rule (5) of the SHB simplifies to

$$x_{k+1} = x_k - \omega \frac{\mathbf{A}_{i:}x_k - b_i}{\|\mathbf{A}_{i:}\|_2^2}\mathbf{A}_{i:}^\top + \beta(x_k - x_{k-1}).$$

This is a randomized Kaczmarz method (RK) with momentum. Note that for $\beta = 0$ and $\omega = 1$ this reduces to the celebrated *Randomized Kaczmarz method* (RK) of Strohmer and Vershynin [13]. In Figure 1, RK with momentum is tested for several values of the momentum parameters β and fixed stepsize $\omega = 1$. For the evaluation we use both the relative error measure $\|x_k - x_*\|^2/\|x_0 - x_*\|^2$ and the function suboptimality $f(x_k) - f(x_*)$. The starting point is chosen as $x_0 = 0$. For the horizontal axis we use either the number of iterations or the wall-clock time measured using the tic-toc Julia function. It is clear that in this setting the addition of momentum parameter is beneficial and leads to faster convergence.

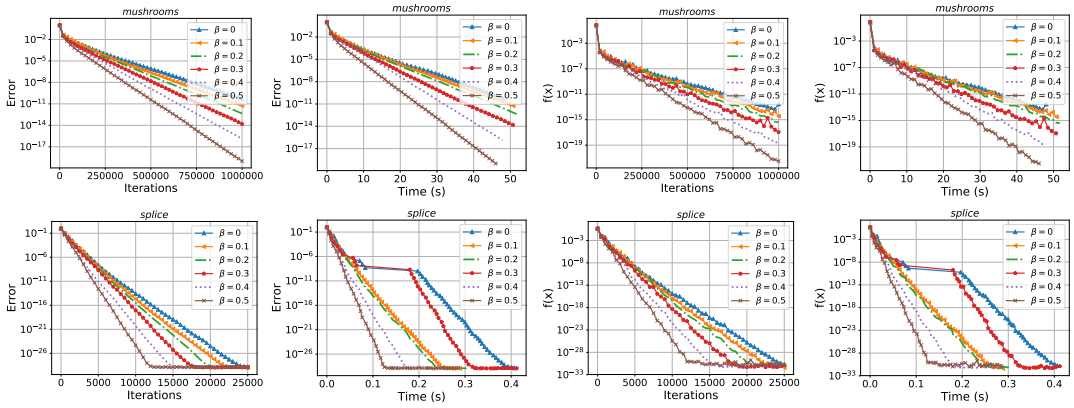

Figure 1: The performance of RK with momentum for several momentum parameters β on real data from LIBSVM [1]. mushrooms: $(n, d) = (8124, 112)$, splice: $(n, d) = (1000, 60)$. The graphs in the first (second) column plot iterations (time) against residual error while those in the third (forth) column plot iterations (time) against function values. The "Error" on the vertical axis represents the relative error $\frac{\|x_k - x_*\|^2}{\|x_*\|^2}$ and the function values $f(x_k)$ refer to function (1).

References

[1] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[2] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

[3] S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. *arXiv:1609.04228*, 2016.

[4] R.M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM. J. Matrix Anal. & Appl.*, 36(4):1660–1690, 2015.

[5] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

[6] J. Konečný, J. Liu, , P. Richtárik, and M. Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2016.

[7] J. Konečný and P. Richtárik. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 3(9):1–14, 2017.

[8] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

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

[10] B.T. Polyak. Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*, 1987.

[11] P. Richtárik and M. Takáč. Stochastic reformulations of linear systems: algorithms and convergence theory. *arXiv:1706.01108*, 2017.

[12] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

[13] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.

[14] I. Sutskever, J. Martens, G.E. Dahl, and G.E. Hinton. On the importance of initialization and momentum in deep learning. *ICML (3)*, 28:1139–1147, 2013.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[16] T. Yang, Q. Lin, and Z. Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.