Computational Optimization and Applications (2021) 80:681–685
https://doi.org/10.1007/s10589-021-00327-x

COAP 2020 best paper prize

Each year, the editorial board of Computational Optimization and Applications selects a paper from the preceding year's publications for the Best Paper Award. This article highlights the research related to the award winning paper of Nicolas Loizou (Johns Hopkins University) and Peter Richtárik (King Abdullah University of Science and Technology) whose award-winning paper "Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods" was published in volume 77, pages 653–710.

Their paper [10] studies several classes of stochastic optimization algorithms enriched with the *heavy ball momentum*. Among the methods studied are: stochastic gradient descent (SGD), stochastic Newton (SN), stochastic proximal point (SPP) and stochastic dual subspace ascent (SDSA). This was the first time momentum variants of several of these methods were studied.

The baseline first-order method for minimizing a differentiable function f is the *gradient descent (GD)* method, $x^{k+1} = x^k - \omega \nabla f(x^k)$, where $\omega > 0$ is a stepsize [3]. For μ-strongly convex function f with L-Lipschitz gradient, it is well-known that GD converges to the solution with the linear rate $\mathcal{O}((L/\mu)\log(1/\epsilon))$ [12]. To improve the convergence behavior of GD, Boris Polyak in a seminal work [14, 15] proposed to modify the update rule by introducing a (heavy ball) momentum term, $\beta(x^k - x^{k-1})$, where $\beta > 0$ is a momentum parameter. This leads to the GD method with momentum, popularly known as the *heavy ball method:* $x^{k+1} = x^k - \omega \nabla f(x^k) + \beta(x^k - x^{k-1})$. More specifically, Polyak proved that with the correct choice of the stepsize ω and the momentum parameter β, a local accelerated linear convergence rate of $\mathcal{O}(\sqrt{L/\mu}\log(1/\epsilon))$ can be achieved in the case of twice continuously differentiable μ-strongly convex objective functions f with L-Lipschitz gradient [14, 15]. Since its original inception, the optimization community has focused on studying the properties of the heavy ball method in several settings [4, 7, 13].

The stochastic variant of the algorithm, the stochastic heavy ball method (also known as SGD with momentum), where only an unbiased estimator $g(x^k)$ of the true gradient $\nabla f(x^k)$ is used in each step,

$$x^{k+1} = x^k - \omega g(x^k) + \beta(x^k - x^{k-1}),$$

has been immensely popular in the machine learning community as it helps to speed up the training of modern machine learning models [6, 19, 20]. In these scenarios, it has been observed that the use of momentum on top of stochastic algorithms can

significantly improve the training time and quality of the trained model. However, despite the popularity of the method and the considerable amount of work focusing on understanding its properties, the convergence behavior of the *stochastic* variants of the algorithm was not understood well.

In their paper [10], Loizou and Richtárik focus precisely on this and provide a robust theoretical analysis and understanding of how the heavy ball momentum interacts with the update rules of several popular stochastic optimization algorithms. Their convergence analysis focuses on solving large-scale convex quadratic problems where all methods under study (SGD, SN, SPP and SDSA) are equivalent [16]. In particular, they prove global non-asymptotic linear convergence rates for all these stochastic methods and for various measures of success, including primal function values, primal iterates, and dual function values. This seems to be the first paper providing the analysis for momentum variants of SN, SPP and SDSA.

Loizou and Richtárik [10] provide several (global and non-asymptotic) linear convergence results for the primal methods SGD/SN/SPP with momentum. A linear rate for the decay of the expected squared distance to the solution, $\mathbb{E}\left[\|x^k - x^*\|_{\mathbf{B}}^2\right]$, where $\mathbf{B}$ is a positive definite matrix defining the norm, was established as well, for a range of stepsizes $\omega > 0$ and momentum parameters $\beta \geq 0$. The same rate was proved to hold for i) the decay of the expected function suboptimality $\mathbb{E}\left[f(x_k)\right] - f(x_*)$ of the stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) := \mathbb{E}\left[f_{\mathbf{S}}(x)\right],$$

where $\mathbf{S}$ is a random matrix defining the source of randomness, and for ii) the convergence of the dual objective to the optimum in case of SDSA with momentum. No linear rates for any of these methods with momentum were known before.

Loizou and Richtárik [10] further study the decay of the larger quantity $\|\mathbb{E}\left[x^k - x^*\right]\|_{\mathbf{B}}^2$ to zero. In this case, the authors established an *accelerated* linear rate, which depends on the square root of the condition number. This is a quadratic speedup when compared to the no-momentum methods as these depend on the condition number. This is the first time an accelerated rate is obtained for the stochastic heavy ball method (mSGD). Prior to their work, no global non-asymptotic accelerated linear rates were established even in the non-stochastic setting (i.e., for the heavy ball method). In addition, under somewhat weaker conditions, the sublinear convergence rate $O(1/k)$ of all primal momentum methods was proved for the Cesàro averages of the iterates and for $\mathbb{E}\left[f(\hat{x}_k)\right] - f(x_*)$ (here $\hat{x}_k = \frac{1}{k}\sum_{t=0}^{k-1} x_t$).

Moreover, Loizou and Richtárik [10] propose a novel concept, for which they coined the name *stochastic momentum*, aimed at decreasing the cost of performing the momentum step. Stochastic momentum is a stochastic (coordinate-wise) approximation of the deterministic momentum and hence is much less costly, which in some situations leads to computational savings in each iteration. The authors analyze the SGD, SN, and SPP methods with stochastic momentum and prove linear convergence rates. They also show that in some sparse data regimes, the overall complexity of SGD with stochastic momentum is better than the overall complexity of SGD with the classical deterministic momentum.

As explained in their work, all proposed algorithms (with momentum or stochastic momentum) can be interpreted as sketch-and-project methods [5] for a solving consistent linear system. This interpretation of the algorithms allows for the recovery of a comprehensive array of well-known methods as special cases by a careful choice of the parameters of the algorithms. To this end, [10] was the first paper to analyze momentum variants of several popular algorithms for solving large-scale linear systems, including the randomized Kaczmarz method [18], randomized coordinate descent [8], Gaussian Kaczmarz [5], and their block variants.

Extensive numerical testing on artificial and real datasets, including data coming from average consensus problems, has also been presented in [10] to corroborate their theoretical results, and to demonstrate the practical benefits of adding the momentum term. More specifically, the authors evaluate the performance of the randomized Kaczmarz method with momentum and the randomized coordinate descent method with momentum for solving both synthetic consistent Gaussian systems and consistent linear systems with real matrices. It was also experimentally shown that the addition of the momentum accelerates the pairwise randomized gossip algorithm for solving the average consensus problem [1].

Since the publication of this paper in Computational Optimization and Applications, the proposed proof techniques and ideas have already served as a starting point for the analysis of SGD with momentum and the development of SN, SPP and SDSA methods with momentum for more general problem classes (beyond special quadratics), including convex and also non-convex optimization problems [2, 9, 11, 17].

Finally, this work provides a bridge across several communities, including numerical linear algebra, stochastic optimization, machine learning, computational geometry, fixed point theory, applied mathematics and probability theory.

Nicolas Loizou is an Assistant Professor in the Department of Applied Mathematics and Statistics and the Mathematical Institute for Data Science (MINDS) at The Johns Hopkins University. Prior to this, he was an IVADO postdoctoral research fellow at Mila - Quebec Artificial Intelligence Institute and Université de Montréal. He received his Ph.D. in Optimization and Operational Research from the University of Edinburgh, United Kingdom, in 2019.

Peter Richtárik is a Professor of Computer Science at the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. Prior to joining KAUST, he was an Associate Professor of Mathematics at the University of Edinburgh, and held postdoctoral and visiting positions at Université Catholique de Louvain, Belgium, and the University of California, Berkeley, USA, respectively. He received his PhD in 2007 from Cornell University, USA.

References

1. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Randomized gossip algorithms. IEEE Trans. Inf. Theory **14**(SI), 2508–2530 (2006)
2. Can, B., Gurbuzbalaban, M., Zhu, L.: Accelerated linear convergence of stochastic momentum methods in wasserstein distances. In ICML, pages 891–901 (2019)
3. Cauchy, A.: Méthode générale pour la résolution des systemes d'équations simultanées. Comp. Rend. Sci. Paris **25**(1847), 536–538 (1847)
4. Ghadimi, E., Feyzmahdavian, H.R., Johansson, M.: Global convergence of the heavy-ball method for convex optimization. In ECC, pages 310–315. IEEE (2015)
5. Gower, R.M., Richtárik, P.: Randomized iterative methods for linear systems. SIAM J. Matrix Anal. Appl. **36**(4), 1660–1690 (2015)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In NeurIPS, pages 1097–1105, (2012)
7. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim. **26**(1), 57–95 (2016)
8. Leventhal, D., Lewis, A.S.: Randomized methods for linear constraints: convergence rates and conditioning. Math. Oper. Res. **35**(3), 641–654 (2010)
9. Liu, Y., Gao, Y., Yin, W.: An improved analysis of stochastic gradient descent with momentum. NeurIPS 33 (2020)
10. Loizou, Nicolas, Richtárik, Peter: Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. Comput. Optim. Appl. **77**(3), 653–710 (2020)

11. Loizou, N., Richtárik, P.: Revisiting randomized gossip algorithms: General framework, convergence rates and novel block and accelerated protocols. IEEE Trans. Inf. Theory (2021)
12. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer (2013)
13. Ochs, P., Brox, T., Pock, T.: iPiasco: Inertial proximal algorithm for strongly convex optimization. J. Math. Imaging Vis. **53**(2), 171–181 (2015)
14. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
15. Polyak, B.T.: Introduction to optimization. translations series in mathematics and engineering. Optim. Softw. (1987)
16. Richtárik, Peter, Takác, Martin: Stochastic reformulations of linear systems: algorithms and convergence theory. SIAM J. Matrix Anal. Appl. **41**(2), 487–524 (2020)
17. Sebbouh, O., Gower, R.M., Defazio, A.: Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In COLT, pages 3935–3971. PMLR (2021)
18. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl. **15**(2), 262–278 (2009)
19. Sutskever, I., Martens, J., Dahl, G.E., Hinton, G.E.: On the importance of initialization and momentum in deep learning. ICML **28**, 1139–1147 (2013)
20. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In CVPR, pages 1–9 (2015)