The Role of Without-Replacement Sampling in Least Square Problems and Additive Convex Optimization: New Results and Algorithms

Mert Gürbüzbalaban, Asu Ozdaglar, Pablo Parrilo

Abstract

Consider the over-determined least square problem where

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{2}\|Ax-b\|^2, \tag{1}$$

$b=(b_1,b_2,\ldots,b_m)^T$ is an m-dimensional column vector and A is a real $m\times n$ matrix with row vectors a_i for $i=1,2,\ldots,m$. We focus on the case when m is large, which arises for instance in machine learning and big data analysis applications, see e.g. [4]. The randomized Kaczmarz (RK) method is an iterative algorithm for solving this problem whose rate does not depend on m [24], therefore when m is large, it often converges faster than other standard methods such as conjugate gradient or GMRES [12, 24, 16]. For the simplicity of the presentation of the RK method and our results in the discussion below, we assume that the rows of A are normalized, i.e.

$$\|a_i\|=1, \quad i=1,2,...,m,$$

and A is of full rank. Note that the full rankness of A guarantees that $f(x)$ is strongly convex and a solution x^* to the least squares problem (1) exists and is unique. At every iteration $k\geq 0$, the RK method selects a row a_{i_k} of the A matrix and computes

$$x^{k+1}=x^k-\alpha_k(a_{i_k}x^k-b_{i_k})a_{i_k}^T, \quad k\geq 0, \tag{2}$$

where i_k is sampled independently and uniformly *with replacement* from the set of indices $\{1,2,\ldots,m\}$, $\alpha_k>0$ is denoted as the *relaxation parameter* and $x^0\in\mathbb{R}^n$ is a given starting point. This algorithm was originally proposed in [8] with $\alpha_k=1$ for solving consistent linear systems where the row indices i_k are selected in a deterministic and cyclic fashion from $\{1,2,\ldots,m\}$, i.e.

$$i_k = k \pmod{m} + 1.$$

However, the original algorithm may converge slowly in some cases, for instance, when many neighboring rows are identical [12]. Therefore, alternative choices of α_k and choosing i_k in a random fashion instead of a cyclic fashion have been proposed in the literature to accelerate convergence [24, 12, 3, 5, 14]. In particular, for inconsistent systems, $\alpha_k\to 0$ is required in (2) for obtaining global convergence to the solution of (1) and a possible choice would be $\alpha_k=\mathcal{O}(1/k)$. A key observation to emphasize here is the fact that the RK method is a special case of the *stochastic gradient method* [17]. To see this, we can rewrite (1) and (2) as

$$\min_{x\in\mathbb{R}^n} f(x) \equiv \sum_{i=1}^m f_i(x) \quad \text{where} \quad f_i(x)=(a_ix-b_i)^2 \tag{3}$$

$$x^{k+1} = x^k-\alpha_k\nabla f_{i_k}(x^k), \quad k\geq 0, \tag{4}$$

where $f_i(x)$ are called *the component functions* whose sum is the objective $f(x)$ and the problem (3) is a particular instance of the *additive convex cost optimization problems* where the component

functions are quadratics. The iterations defined by (4) coincide exactly with the stochastic gradient iterations applied to $f(x)$ with step size α_k [23, 1, 17]. This allows us to leverage some of the favorable iteration complexity results developed for the stochastic gradient (SG) method to understand the convergence properties of the RK method. In particular, for the strongly convex problem (3), it is known that SG has a lower bound of $\Omega(1/k)$ [19] and this lower bound can be achieved (up to a constant factor) by choosing the step size as $\alpha_k = R/k$ for any $R \geq 1/\mu$ where μ is the strong convexity constant of f [7]. Then, with this choice of the step size, the suboptimality in objective values decays with $f(x_k) - f(x_*) = \mathcal{O}(1/k)$, which is optimal (up to a constant factor). Another possible choice of step size to achieve the lower bound is to use a slower decaying step size of the form $\alpha_k = \mathcal{O}(1/k^s)$ for $s \in (1/2, 1)$ and averaging the iterates. This procedure is known as the Polyak-Ruppert averaging and its advantage compared to the choice of $\alpha_k = R/k$ is that it does not require the knowledge (or accurate estimation) of the strong convexity constant of f [15, 9].

Although SG and its performance under different step size rules have been well studied in the literature, an abundance of numerical experiments with SG has revealed a curious phenomenon about its convergence properties that is not well understood: If the order is sampled uniformly from $\{1, 2, \ldots, m\}$ *without replacement* instead of with replacement in SG or in RK, then the convergence is often much faster, obeying an empirical $\mathcal{O}(1/k^2)$ convergence rate compared to $\Omega(1/k)$ convergence of with-replacement sampling [2, 22]. However, understanding this discrepancy in convergence rate between the with- and without-replacement sampling for the SG method (and also for the RK method) has been a long-standing open problem [21, 1, 18].

Recently, in [7], we answer this question by showing that SG (and also RK) with iterate averaging and a diminishing step size $\alpha_k = \Theta(1/k^s)$ for $s \in (1/2, 1)$ converges at rate $\Theta(1/k^{2s})$ with probability one in the suboptimality of the objective value, thus improving upon the $\Omega(1/k)$ rate in expectation of the with-replacement sampling. Furthermore, our results are applicable to more than just the linear least squares problems of the form given by (1) as long as the component functions are twice continuously differentiable and the sum function $f(x)$ is strongly convex. For example, our results apply to regularized non-linear least squares problems or regularized logistic regression problems.

Our analysis draws on the theory of Polyak-Ruppert averaging and relies on viewing SG as a gradient descent method with gradient errors that are dependent to each other among the iterations (the dependency is a consequence of without-replacement sampling). We first decouple the gradient errors into an independent term and another term dominated by α_k^2. This allows us to apply the law of large numbers to an appropriately weighted version of the gradient errors, where the weights depend on the step size. We also provide high probability convergence rate estimates that show decay rates of different terms and allow us to propose a novel variant of the RK algorithm (and also of the SG algorithm) with convergence rate $\mathcal{O}(1/k^2)$. Finally, we show that the $\mathcal{O}(1/k^2)$ rate can also be achieved in expectation for the $s = 1$ case by appropriately adjusting the step size with respect to the strong convexity constant of the objective. Fundamental to the analysis in [7] is the new convergence rate results obtained in [6] for the deterministic incremental gradient method with an arbitrary deterministic order.

We also note that the randomized coordinate descent (RCD) method is another related effective algorithm for solving large-scale least squares and optimization problems [20, 12]. An interesting connection between RK and RCD is that for consistent linear systems $Ax = b$, the RK update (2) is equivalent to one step of the RCD method applied to the dual problem

$$\min_y \frac{1}{2}\|A^T y\|^2 - b^T y, \tag{5}$$

where the update is taken with respect to the i_k-th coordinate with step size α_k (where the primal variables x and duals y are related through $x = A^T y$) [12, 11]. Interestingly, RCD is also equivalent to applying a randomized Gauss-Seidel algorithm to the system $Ax = b$ [11, 13]. Then, a natural question that arises is how much potential speed-up without-replacement sampling can offer compared to with-replacement sampling in the context of RCD, a key question that has been open in its full generality with direct implications to randomized Gauss-Seidel methods. Recently, Lee and Wright [10] constructed an example of a symmetric matrix A where applying RCD to the dual problem (5) using without-replacement sampling accelerates the convergence asymptotically by a factor of 2 compared to with-replacement sampling. Motivated by this work, for every positive integer n, we construct a family of novel examples consisting of $2n \times 2n$ symmetric matrices A_n where without-replacement sampling results in a faster convergence with respect to with-replacement sampling by a factor of 2. Our example sheds light onto the behavior of with-replacement based algorithms and their acceleration potential. A key point in the construction of our example is to exploit the connections between RK, RCD and the randomized Gauss-Seidel methods.

Our new theoretical results for RK, SG and RCD methods highlight the potential of without-replacement sampling in randomized algorithms for numerical linear algebra and optimization. Furthermore, our results have yielded novel and improved variants of the RK and SG methods with significantly faster convergence rates in theory and in practice. We also present numerical experiments demonstrating our theoretical results and the effectiveness of our proposed algorithms.

References

[1] D. Bertsekas. *Nonlinear programming.* Athena Scientific, 1999.

[2] L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the Symposium on Learning and Data Science, Paris*, 2009.

[3] L. Dai, M. Soltanalian, and K. Pelckmans. On the randomized Kaczmarz algorithm. *IEEE Signal Processing Letters*, 21(3):330–333, March 2014.

[4] A. Défossez and F. R Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *AISTATS*, 2015.

[5] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[6] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Convergence Rate of Incremental Gradient and Newton Methods. *arXiv preprint arXiv:1510.08562*, October 2015.

[7] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Why random reshuffling beats stochastic gradient descent. *arXiv preprint arXiv:1510.08560*, 2015.

[8] Stefan Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.

[9] H. J. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[10] Ching-Pei Lee and Stephen J Wright. Random permutations fix a worst case for cyclic coordinate descent. *arXiv preprint arXiv:1607.08320*, 2016.

[11] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.

[12] Ji Liu and Stephen Wright. An accelerated randomized kaczmarz algorithm. *Mathematics of Computation*, 85dd(297):153–178, 2016.

[13] Anna Ma, Deanna Needell, and Aaditya Ramdas. Convergence properties of the randomized extended gauss–seidel and kaczmarz methods. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1590–1604, 2015.

[14] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[15] E. Moulines and F. R. Bach. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *Advances in Neural Information Processing*, pages 451–459, 2011.

[16] D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.

[17] D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2014.

[18] Deanna Needell and Joel A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra and its Applications*, 441:199 – 221, 2014.

[19] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009.

[20] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[21] B. Recht and C. Ré. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. *JMLR Workshop and Conference Proceedings*, 23:11.1–11.24, 2012.

[22] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[23] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

[24] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

Go back to Thursday's schedule.   *Go back to Parallel Talks #6.*   *Go back to speaker index.*