RANDOMIZED COORDINATE DECENT ALGORITHMS

ANA BOERIU, JING LYU, AND WENZHUO WU

Abstract. In this report, we focus on the most essential variants of coordinate descent algorithm: randomized coordinate descent and accelerated randomized coordinate descent algorithms. The fundamental theorems behind the algorithms are introduced with reference to convex objectives. Besides, we compare the iteration complexity of randomized coordinate descent and accelerated randomized coordinate descent for Kaczmarz algorithm applied in linear equations and unconstrained minimization problem. The simulation shows consistent results with the theoretical methodology: accelerated randomized CD algorithm performs best when solving linear system and empirical risk minimization problem.

1. Introduction

Coordinate Decent (CD) is an optimization algorithm that solves optimization problems by successively performing approximate minimization along coordinate directions or coordinate hyperplane [Ste15]. CD has many applications including lasso, and linear regression. One of the applications of CD is solving linear equations using the Kaczmarz Algorithm. This algorithm is an iterative algorithm that is used for solving linear systems $Aw = b$ [Kac37]. First discovered by a Polish mathematician, Stefan Kaczmarz, this algorithm is a sequential projection action method where only one row is used in each iteration [YCJ09]. Because of its simplicity, the Kaczmarz method has been used in numerous applications such as image reconstruction and signal processing.

Build as an extension to CD, randomized CD is an optimization algorithm that was applied to the problem of minimizing a smoothing function. The CD method can be deterministic or randomized depending on the choice of the update coordinates. If the coordinate indices i_k is sampled uniformly from the set $1, 2, ..., n$, the resulting method is called the randomized coordinate descent method. Randomized CD approach optimization problems by performing gradient descent along the subgroups of coordinates. Additionally, randomized CD algorithm can be transformed in a computationally efficient algorithm applied by an appropriate multi-step strategy known as accelerated randomized coordinate descent algorithm. It is widely used in optimization due to its efficiency and scalability to large-scale problems. Lee and Sidford [LS13] found that it is possible to implement the accelerated randomized

CD approach efficiently for linear systems $Aw = b$ and empirical risk minimization problems.

In this project we focus on the most important variants of coordinate descent algorithms: randomized CD and accelerated randomized CD algorithms. We apply those methods to the Kaczmarz algorithm in linear system and convex and smooth functions to explore their properties. This project provides an in depth summary of statistical algorithms. By the end of this report the reader will be able to understand the methodology and application of randomized CD algorithm and its accelerated version.

2. Algorithm Description

In this project, our focus is the unconstrained minimization problem, which is:

$$\min_x f(x) \tag{1}$$

where f is a convex and smooth function. We assume strong convexity with respect to the Euclidean norm where there exists $\sigma > 0$ such that $f(y) \geq f(x) + \nabla f(x)^T + \frac{\sigma}{2}\|y - x\|_2^2$, for all x, y. We also define the gradient of function f is coordinate-wise Lipschitz continuous with constants $L_i = L_i(f)$ if $\|\nabla f_i(x + U_i h_i) - \nabla f_i(x)\|_{(i)}^* \leq L_i \|h_i\|_{(i)}, h_i \in \mathbb{R}^{n_i}, i = 1, ..., n, x \in \mathbb{R}^n$. Algorithm 1 shows the randomized coordinate descent framework for continuously differentiable minimization. In each iteration, i_k is randomly selected from {1,2,3,...,n} with equal probability and corresponding gradient ∇f at the present point will be evaluated. Then the i_k component of x will be followed by an adjustment in the opposite direction to current gradient component. The current selection of i_k is independent of the previous selections.

Algorithm 1 Randomized Coordinate Descent for (1)

Choose $x^0 \in \mathbb{R}^n$;
Set $k \leftarrow 0$;
repeat
  Choose index i_k with uniform probability from {1,2,3,...,n} independently of choices at prior iterations;
  Set $x^{k+1} \leftarrow x^k - \alpha_k [\nabla f(x^k)]_{i_k} e_{i_k}$ for some $\alpha_k > 0$;
  $k \leftarrow k + 1$;
until termination test satisfied;

The accelerated randomized coordinated descent algorithm (Algorithm 2) is proposed by Nesterov [Nes12]. The assumption of this algorithm is that there is an

estimate strong convexity $\sigma \geq 0$ and an estimate of the component-wise L_i. It requires manipulation of the generally dense vector y and v. Additionally, the gradient changes of y^k is more extensively for iterations. However, it is efficient for specific problems such as linear system and empirical risk minimization problems.

Algorithm 2 Accelerated Randomized Coordinate Descent for (1)

Choose $x^0 \in \mathbb{R}^n$;
Set $k \leftarrow 0, v^0 \leftarrow 0, \gamma_{-1} \leftarrow 0$;
repeat
  Choose γ_k to be the larger root of $\gamma_k^2 - \frac{\gamma_k}{n} = (1 - \frac{\gamma_k \sigma}{n})\gamma_{k-1}^2$.
  Set $\alpha_k \leftarrow \frac{n - \gamma_k \sigma}{\gamma_k (n^2 - \sigma)}, \beta_k \leftarrow 1 - \frac{\gamma_k \sigma}{n}$;
  Set $y^k \leftarrow \alpha_k v^k + (1 - \alpha_k) x^k$;
  Choose index i_k with uniform probability from {1,2,3,...,n} and set $d^k = [\nabla f(y^k)]_{i_k} e_{i_k}$;
  Set $x^{k+1} \leftarrow y^k - (1/L_{i_k}) d^k$;
  Set $v^{k+1} \leftarrow \beta_k v^k + (1 - \beta^k) y^k - (\gamma_k / L_{i_k} d^k)$;
  $k \leftarrow k + 1$
until termination test satisfied;

Assuming step size is 1, the iteration of randomized Kaczmarz algorithm by applying algorithm 1 is

$$w^{k+1} \leftarrow w^k - (A_{ik} A^T x^k - b_{ik}) A_{ik}^T = w^k - (A_{ik} w^k - b_{ik}) A_{ik}^T).$$

The iteration of accelerated randomized Kaczmarz algorithm by applying algorithm 2 is

$$w^{k+1} \leftarrow y^k - (A_{ik} y^k - b_{ik}) A_{ik}^T \text{ where } y^k \leftarrow \alpha_k v^k + (1 - \alpha_k) w^k$$

3. Main Results

The following assumption leads to the main theorems of the paper.

Assumption 1. f in (1) is convex and uniformly Lipschitz continuously differentiable, and it reaches its minimum value f^* on a set S. The level set for f defined by x^0 is bounded around a finite R_0,

$$\max_{x^* \in S} \max_x \{\|x - x^*\| : f(x) \leq f(x^0)\} \leq R_0 \tag{2}$$

The most rigorous part of the assumption is the initial point should be close to the true value so that $f(x^0)$ is not larger than $f(x)$ too much and the level set it bounded as desired. It also implicitly requires the domain set S is not very large. Larger R_0 will cause more iterations.

Theorem 1. Suppose assumption 1 holds and $\alpha \equiv 1/L_{max}, L_{max} = \max_{i=1,2,\dots,n} L_i$, for $k > 0$, we have

$$E(f(x^k)) - f^* \leq \frac{2nL_{max}R_0^2}{k} \tag{3}$$

Specifically, when $\sigma > 0$

$$E(f(x^k)) - f^* \leq (1 - \frac{\sigma}{nL_{max}})^k (f(x^0) - f^*) \tag{4}$$

Randomized Kaczmarz algorithm is a special case of theorem 1. In terms of the system of linear equations $Aw = b$, where $A \in \mathbb{R}^{m\times n}$ and Lagrangian dual is $\min_{x\in\mathbb{R}^n} \ f(x) := \frac{1}{2}\|A^T x\|_2^2 - b^T x$, where $w = A^T x$. The iteration of randomized Kaczmarz algorithm by applying algorithm 1 is $w^{k+1} \leftarrow w^k - (A_{ik}A^T x^k - b_{ik})A_{ik}^T = w^k - (A_{ik}w^k - b_{ik})A_{ik}^T$) [SV09]. Denoting $\lambda_{min,nz}$ as the minimum nonzero eigenvalue of AA^T and $P(\cdot)$ as the projection onto the solution set of $Aw = b$, we have

$$E\|w^k - P(w^k)\|_2^2 \leq \left(1 - \frac{\lambda_{min,nz}}{m}\right)^k \|w^0 - P(w^0)\|_2^2 \tag{5}$$

The following convergence result of Algorithm 2 proved by Nesterov [Nes12] implies a reduction of computational complexity to meet a specified error tolerance.

Theorem 2. Suppose assumption 1 holds and define

$$S_0 := \sup_{x^*\in S} L_{max}\|x^0 - x^*\|_2^2 + (f(x^0) - f^*)/n^2$$

Then for all $k \geq 0$ we have

$$\begin{aligned} &E(f(x^k)) - f^* \\ &\leq S_0 \frac{\sigma}{L_{max}} \left[\left(1 + \frac{\sqrt{\sigma/L_{max}}}{2n}\right)^{k+1} - \left(1 - \frac{\sqrt{\sigma/L_{max}}}{2n}\right)^{k+1}\right]^{-2} \leq S_0 \left(\frac{n}{k+1}\right)^2 \end{aligned} \tag{6}$$

Based on Theorem 2, in the strong convex case $\sigma > 0$, the convergence rate of (6) is significantly faster than the rate (4) for Algorithm 1.

The number of iterations required to meet a specified error tolerance of different randomized CD variants was compared. The iteration complexity of each type of randomized CD in unconstrained case is shown in Table 1.

4. Key Proof Ideas

Theorem 1 [Ste15]**:** The proof first shows the deduction:

$$f(x^{k+1}) - f(x^k) < -[\nabla f(x^k)]_{i_k}^2 / 2L_{max}$$

Algorithm	Assumption	Iteration Complexity
Randomized CD	Assumption 1	$\mathcal{O}(1/\epsilon)$
	Assumption 1, strong convexity	$\mathcal{O}(log(1/\epsilon))$
Randomized Kaczmarz		$\mathcal{O}(log(1/\epsilon))$
Accelerated Randomized CD	Assumption 1	$\mathcal{O}(1/\sqrt{\epsilon})$
	Assumption 1, strong convexity	$\mathcal{O}(1/\sqrt{\epsilon})$
Accelerated Randomized Kaczmarz		$\mathcal{O}(log(1/\epsilon))$

Table 1. Iteration complexity for different randomized CD variants

based on Taylor expansion and the Lipschitz assumptions:

$$[\nabla^2 f(x^k)]_{ii} < L_i < L_{max}.$$

Then, using the convexity of f to bound the difference $f(x^k) - f^*$ and the bounded level set assumption to replace $\|x^k - x^*\|$ by the constant R_0. Finally, the expectation is applied on both sides and the first bound in the theorem is found. For the strongly convex f, the proof follows the same steps using its strong convexity assumption and leads to faster convergence rate. A similar result is proved for l_1-regularized problems [SST11] .

Theorem 2 [Nes12]**:** The proof starts with bounding the distance between v_k and the optima point. Two additional parameters are introduced to measure the growth rate. Then using the relationship among the denoted parameters and the technique of applying the expectation w.r.t r_{i_k} on both sides of the inequality, the upper bound of the difference between the expected value of current step and the minimum is obtained. The growth of the two coefficients is estimated. Using the trivial inequality, the final inequality format is proved by induction. In the strong convexity situation, the $(1 + \frac{\sqrt{\sigma/L_{max}}}{2n})^{k+1}$ would dominate and the convergence rate of accelerated algorithm is much faster than that of the regular version.

Linear Convergence Rate for Randomized Kaczmarz Algorithm [SV09]**:** In terms of the proof of the linear convergence for Kaczmarz iteration, it shows after we normalized the rows, it has

$$\|w^{k+1} - P(w^{k+1})\|^2 \leq \|w^k - A_{ik}w^k - b_{ik} - P(w^k)\|^2 \leq \frac{1}{2}\|w^k - P(w^k)\|^2 - (A_{ik}w^k - b_{ik})^2$$

By taking expectations of both sides with respect to i_k, we have

$$E_{ik}\|w^{k+1} - P(w^{k+1})\|^2 = \frac{1}{2}\|w^k - P(w^k)\|^2 - \frac{1}{m}\|Aw^k - b)\|^2 \leq \left(1 - \frac{\lambda_{min,nz}}{m}\right)\|w^k - P(w^k)\|^2.$$

By applying recursive application of this formula, (5) can be proved.

5. Full Proof

Proof of Theorem 1 [Ste15]**:** Based on the Taylor expansion, f is uniformly Lipschitz continuously differentiable and the fact that $L_{max} = max_{i=1,..,n} L_i$, we have

$$\begin{aligned} f(x^{k+1}) &= f(x^k - \alpha_k [\nabla f(x^k)]_{ik} e_{ik}) \\ &\leq f(x^k) - \alpha_k \left(1 - \frac{L_{max}}{2}\alpha_k\right) [\nabla f(x^k)]_{ik}^2 \\ &= f(x^k) - \frac{1}{2L_{max}} [\nabla f(x^k)]_{ik}^2 \end{aligned}$$

Then take the expectation of both sides, we have

$$E_{ik} f(x^{k+1}) = f(x^k) - \frac{1}{2nL_{max}} \|\nabla f(x^k)\|^2.$$

Subtracting $f(x^*)$ from both sides and taking expectation w.r.t. all random variables, and denote $\theta_k = E(f(x^k)) - f^*$, we obtain

$$\theta_{k+1} \leq \theta_k - \frac{1}{2nL_{max}} [E(\|\nabla f(x^k)\|]^2 \tag{7}$$

Taking expectation of convexity, we get $E(\|\nabla f(x^k)\|) \geq \frac{1}{R_0}\theta_k^2$, by applying this recursively, we obtain

$$\frac{1}{\theta_k} \geq \frac{1}{\theta_0} + \frac{k}{2nL_{max}R_0^2} \geq \frac{k}{2nL_{max}R_0^2}$$

so (3) holds. In addition, for the strong convex with $\sigma > 0$, taking the minimum of both sides of $f(y) \geq f(x) + \nabla f(x)^T + \frac{\sigma}{2}\|y - x\|_2^2$, and use the bound in (7), we get

$$\theta_{k+1} \leq \left(\theta_k - \frac{\sigma}{nL_{max}}\right) \theta_k$$

By applying recursive application of this formula, (4) can be proved.

Proof of Theorem 2 [Nes12]**:** Denote $\xi_k = \{i_0, ..., i_k\}, r_k^2 = \|v_k - x^*\|^2$. The algorithm shows $v_k = y_k + \frac{1-\alpha_k}{\alpha_k}(y_k - x_k)$
then we can have

$$\begin{aligned} r_{k+1}^2 = \|v_{k+1}^2 - x^*\|^2 &= \|\beta_k v_k + (1-\beta_k)y_k - x^*\|^2 + \frac{\gamma_k^2}{L_{i_k}}(d_{i_k}^k)^2 + 2\gamma_k\langle d^k, (x^* - \beta_k v_k - (1-\beta_k)y_k)\rangle \\ &\leq \|\beta_k v_k + (1-\beta_k)y_k - x^*\|^2 + L_{max}\gamma_k^2(f(y_k) - f(x_{k+1})) \\ &+ 2\gamma_k\langle d^k, (x^* - y_k + \frac{\beta_k(1-\alpha_k)}{\alpha_k})(x_k - y_k)\rangle \end{aligned}$$

Taking the expectation of both sides w.r.t i_k and using the fact that $\gamma_k^2 - \frac{\gamma_k}{n} = \beta_k\gamma_k\frac{1-\alpha_k}{n\alpha_k}$, we have

$$\begin{aligned}E_{i_k}(r_{k+1}^2) &\le \beta_k r_k^2 + (1-\beta_k)\|y_k - x^*\|^2 + L_{max}\gamma_k^2[f(y_k) - E_{i_k}(f(x_{k+1}))] \\ &\quad + 2\frac{\gamma_k}{n}\langle \nabla f(y_k), x^* - y_k + \frac{\beta_k(1-\alpha_k)}{\alpha_k}(x_k - y_k)\rangle \\ &\le \beta_k r_k^2 - L_{max}\gamma_k^2 E_{i_k}(f(x_{k+1})) + 2\frac{\gamma_k}{n}[f^* + \frac{\beta_k(1-\alpha_k)}{\alpha_k}f(x_k)]\end{aligned}$$

Define $b_0 = 2, a_0 = \frac{1}{n}, b_{k+1} = \frac{b_k}{\sqrt{\beta_k}}, a_{k+1} = \gamma_k b_{k+1}$, then we have

$$b_{k+1}^2 = \frac{1}{\beta_k}b_k^2, a_{k+1}^2 = \gamma_k^2 b_{k+1}^2, \gamma_k\frac{\beta_k(1-\alpha_k)}{n\alpha_k} = \frac{a_k^2}{b_{k+1}^2}$$

Multiplying the inequality above by b_{k+1}^2 and taking the expectation of both sides w.r.t ξ_{k-1}, we obtain

$$\begin{aligned}2a_{k+1}^2(E_{\xi_k}f(x_{k+1}) - f^*) + b_{k+1}^2 E_{\xi_k}(r_{k+1}^2) &\le 2a_k^2(E_{\xi_{k-1}}f(x_k) - f^*) + b_k^2 r_k^2 \\ &\le 2a_0^2(f(x_0) - f^*) + b_0^2\|x_0 - x^*\|^2\end{aligned}$$

Further, since $b_k^2 = \beta_k b_{k+1}^2 = (1 - \frac{\sigma}{n}\gamma_k)b_{k+1}^2 = (1 - \frac{\sigma}{n}\frac{a_{k+1}}{b_{k+1}})b_{k+1}^2$ we have $\frac{\sigma}{n}a_{k+1}b_{k+1} \le b_{k+1}^2 - b_k^2 \le 2b_{k+1}(b_{k+1} - b_k)$ and $b_{k+1} \ge b_k + \frac{\sigma}{2n}a_k$
On the other hand, since $\frac{a_{k+1}^2}{b_{k+1}^2} - \frac{a_{k+1}}{nb_{k+1}} = \frac{\beta_k a_k^2}{b_k^2} = \frac{a_k^2}{b_{k+1}^2}$ we have $\frac{1}{n}a_{k+1}b_{k+1} \le a_{k+1}^2 - a_k^2 \le 2a_{k+1}(a_{k+1} - a_k)$ and $a_{k+1} \ge a_k + \frac{1}{2n}b_k$
Denoting $Q_1 = 1 + \frac{\sqrt{\sigma/L_{max}}}{2n}, Q_2 = 1 - \frac{\sqrt{\sigma/L_{max}}}{2n}$, by induction, we can prove

$$a_k \ge \frac{1}{\sqrt{\sigma}}[Q_1^{k+1} - Q_2^{k+1}], b_k \ge Q_1^{k+1} + Q_2^{k+1}$$

Conclusively, using the inequality $(1+t)^k - (1-t)^k \ge 2kt, t \ge 0$, we have

$$Q_1^{k+1} - Q_2^{k+1} \ge \frac{k+1}{n}\sqrt{\sigma}$$

which gives inequality of the theorem.

6. Simulations

To compare randomized CD and accelerated randomized CD algorithms, we consider the optimization problem in linear system and convex function respectively.

In linear system, we first assume A is a 30×50 real matrix with normalized rows and elements sampled from standard normal distribution. $b = A\mathbb{1}$, where $\mathbb{1}$ is the

vector of ones in $\mathbb{R}^{50}$. Based on the construction of A and b, the solution of the linear equation $AW = b$ exists and is not unique.

For convex function, we consider linear least-squares problem, one of the empirical risk minimization problems. We assume X is a 500×100 real matrix with elements sampled from standard normal distribution. $Y = Xc$, where c is the vector of ones in $\mathbb{R}^{100}$. We are trying to find $\arg\min_{c \in \mathbb{R}^{100}} \|Xc - Y\|_2^2$.

All algorithms stop when $\|x^{k+1} - x^k\|_2 < \epsilon = 10^{-10}$

We start with 100 different initial values for randomized algorithms. The experiment results including average iteration times and error are summarized in Table 2. Standard deviations over 100 different trials are shown in the parenthesis.

	Iteration Times	Error
Randomized Kaczmarz	12447.65(2007.00)	1.34E-05(3.98E-05)
Accelerated Randomized Kaczmarz	7893.61(508.59)	6.02E-09(1.25E-09)
Randomized CD	18311.43(570.56)	4.68E-04(2.55E-03)
Accelerated Randomized CD	7410.82(746.18)	9.78E-07(2.42E-07)

Table 2. Average iteration times over 100 different initial values

Due to the higher cost and larger gradient change of each iteration of accelerated randomized CD compared to the standard method that only requires update of a single component, the accelerated randomized CD maybe not as efficient as the standard method when dealing with certain problems. However, it is possible to apply the accelerated randomized CD method for the linear system problem and empirical risk minimization problem with high efficiency. From the simulations results, we observed the accelerated randomized CD algorithm achieves faster convergence and more accurate results when dealing with linear system and linear least square problem, which is consistent with the theoretical results.

References

[Kac37] S. Kaczmarz, *Angenaherte auflosung von systemen linearer gleichungen*, Bulletin International de l'Academie Polonaise des Sciences et des Lettres **35** (1937), 355–357.

[LS13] Y.T. Lee and A. Sidford, *Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems.*, 54th Annual Symposium on Foundations of Computer Science,pp (2013), 147–156.

[Nes12] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization **22** (2012), 341–362.

[SST11] S. Shalev-Shwartz and A Tewari, *Stochastic methods for l_1-regularized loss minimization.*, Journal of Machine Learning Research **12** (2011), 1865–1892.

[Ste15] J. W. Stephen, *Coordinate descent algorithms.*, Math. Program. **151** (2015), 3–34.

[SV09] T. Strohmer and R. Vershynin, *A randomized kaczmarz algorithm with exponential convergence*, Journal of Fourier Analysis and Applications **15** (2009), 262–278.

[YCJ09] Gabor T. Herman Yair Censor and Ming Jiang, *A note on the behavior of the randomized kaczmarz algorithm of strohmer and vershynin*, Journal of Fourier Analysis and Applications **15** (2009), 431–436.

Appendix A. Additional Details of the Proofs

Provide additional calculations of the proof if needed. You can use unlimited number of pages as appendix.

Email address: aiboeriu@ucdavis.edu

Email address: jjlyu@ucdavis.edu

Email address: wzhwu@ucdavis.edu