Efficiency of coordinate descent methods on huge-scale optimization problems (Nesterov, 2012): a review

Othmane Sebbouh, othmane.sebbouh@gmail.com

***Abstract*— In this report, we will consider the problem of minimizing convex and smooth functions on $\mathbb{R}^N$. To that end, we will use coordinate descent methods. Coordinate descent methods are an alternative to full gradient methods where, at each iteration, a gradient descent step is performed on the function restricted to a small subset of its coordinates, rather than on the whole variable vector. We will consider randomized coordinate descent methods (RCDM). For brievity and clarity, we will only consider unconstrained minimization. We will show that RCDM are very efficient for huge-scale optimization problems and that in some cases, lead to worst-case estimates for convergence rates that can equal those of full gradient methods (FGM), while often guaranteeing cheaper iterations.**

I. INTRODUCTION

Before presenting Nesterov's (2012) main results, we present the simplest versions of coordinate descent methods, which in turn will demonstrate the need for randomized versions of coordinate descent methods.

Consider an componentwise-smooth function f on $\mathbb{R}^N$:

$$\forall i \in \{1, .., N\}, \forall x, h \in \mathbb{R}^N, |\nabla_i f(x + he_i) - \nabla_i f(x)| \leq Lh$$

The most natural idea would then be to chose the coordinate over which the directional descent step is the steepest and perform a 1D-gradient descent. Starting from $x_0 \in \mathbb{R}^N$, this yields the following algorithm:

Algorithm 1 Steepest CD algorithm

For $k = 0, 1, 2...$

- Choose $i_k = \max\limits_{1 \leq i \leq N} |\nabla_i f(x_k)|$
- Update $x_{k+1} = x_k - \frac{1}{L} \nabla_i f(x_k) e_{i_k}$

Then, it can be proven that, with $R \geq \|x_0 - x^*\|$:

$$f(x_k) - f(x^*) \leq \frac{2NMR^2}{k+4}$$

The problem with this algorithm is that the choice of the coordinate in the first step of the algorithm requires the computation of the full gradient, which is costly and misses the main point of using a coordinate descent method.

Another early suggested idea is to cycle through the coordinates and operate a 1D descent for each coordinate:

Algorithm 2 Steepest CD algorithm

For $k = 0, 1, 2...$

- Choose $i_k \equiv k \mod N$
- Update $x_{k+1} = x_k - \frac{1}{L_{i_k}} \nabla_i f(x_k) e_{i_k}$

However, to the best of our knowledge, the best upper boud given for the error in the general case is the one given by Beck and Tetruashvilli (2013):

***Proposition** 1:* Let $\{x_k\}$ be the sequence generated by Algorithm 2. Then:

$$f(x_{k+1}) - f(x^*) \leq 4L_{max}(1 + N^3 \frac{L_{max}^2}{L_{min}^2}) \frac{R^2(x_0)}{k + \frac{8}{N}}$$

where:

- $R^2(x_0) = \max\limits_{x,y \in \mathbb{R}^N} \{\|x - y\| : f(y) \leq f(x) \leq f(x_0)\}$
- $L_{max} = \max\limits_{1 \leq i \leq N} L_i$ and $\min\limits_{1 \leq i \leq N} L_i$

But this bound suffers from a prohibitive dependence on N^3, which clearly poses a problem in huge-scale optimization.

In the article that we analyze, Yurii Nesterov proposes a method that succeeds to overcome the two difficulties related to algorithm 1 and 2: the computational cost of Algorithm 1 and the loose bounds of Algorithm 2. More specifically, he proposes a method where the selection of the coordinates is random, and succeeds to prove tight bounds which, in the worst case, can match with FGM bounds. The advantage of this method is that one iteration can be considerably cheaper than one full gradient descent iteration.

In the next section, we will present Nesterov's main result, which applies to functions which are block coordinate-wise lipshitz. We are going to compare his results with the convergence rates of the classical full gradient method seen in class. In the final section, we will show the benefits of Nesterov's method on a simple function.

II. MAIN RESULTS

A. Notations

Before presenting the notations, note that in this report, to make the notations lighter, we will restrict ourselves to euclidean norms. One major difference between the algorithms presented in the introduction and Nesterov's lies in that the article doesn't consider coordinate descent on a single dimension, but on a block of coordinates, thus yielding a more general result. We note:

- $\mathbb{R}^N = \bigotimes \mathbb{R}^{n_i}, \sum_{i=1}^n n_i = N$
- $I_N = (U_1, ..., U_n), U_i \in \mathbb{R}^{N \times n_i}$
- if a function f is differentiable, its partial gradient in $x^{(i)}$ is noted $f_i^{'}(x)$

Then, if $x = (x^{(1)}, ..., x^{(n)})$, where $x^{(i)} \in \mathbb{R}^{n_i}$, we have: $x = \sum_{i=1}^n U_i x_i$ and $f_i^{'}(x) = U_i^T \nabla f(x) \in \mathbb{R}^{n_i}$

B. Main result

We make the following assumptions:

(A1) f is convex

(A2) f is block-coordinatewise-Lipschitz with constants $(L_i)_{i=1}^n$, i.e.:

$$\forall i = 1, .., N, \forall h_i \in \mathbb{R}^{n_i}, \|f_i^{'}(x + U_i h_i) - f_i{}'(x)\| \leq L_i \|h_i\|$$

Given $\alpha \geq 0$ and $x_0 \in \mathbb{R}^N$, we define the Randomized Coordinate Descent Method as follows:

Algorithm 3 RDCM(α, x_0)

For $k = 0, 1, 2...$

- Choose $i_k \in \{1, ..., n\}$ according to A_α
- Update $x_{k+1} = x_k - \frac{1}{L_{i_k}} U_{i_k} f_{i_k}^{'}(x_k)$

where A_α is a random counter which generates the number $i \in 1, ..., n$ with probability : $p_i^\alpha = \frac{L_i^\alpha}{\sum_{j=1}^n L_j^\alpha}$

Finally, we define $\|x\|_{[\alpha]} = \left[\sum_{i=1}^n L_i^\alpha \|x^{(i)}\|^2\right]^{1/2}$ and $\|g\|_{[\alpha]}^* = \left[\sum_{i=1}^n L_i^{-\alpha} \|g^{(i)}\|^2\right]^{1/2}$.

Note that x_{k+1}, the resulting vector from RCDM(α, x_0), depends on the random variables $\{i_0, ..., i_k\}$. Hence, noting $\xi_k = \{i_0, ..., i_k\}$, we can announce the main theorem:

***Theorem** 1:* Under Assumptions **(A1-2)**: $\forall k \geq 1$,
$$\mathbb{E}_{\eta_{k-1}}[f(x_k)] - f(x^*) \leq \frac{2}{k+4} S_\alpha R_{1-\alpha}^2(x_0)$$
where :

- $R_\beta^2(x_0) = \max_{x,y \in \mathbb{R}^N} \{\|x - y\| : f(y) \leq f(x) \leq f(x_0)\}$
- $S_\alpha = \sum_{i=1}^n L_i^\alpha$

Proof: With $\{x_k\}$ the sequence generated by RDCM(α, x_0), we have :

$$f(x_k) - \mathop{\mathbb{E}}_{i_k}(f(x_{k+1})) = \mathop{\mathbb{E}}_{i_k}[f(x_k) - f(x_{k+1})]$$
$$= \sum_{i=1}^n p_i^\alpha \left[f(x_k) - f(x_k - \frac{1}{L_i} \nabla_i f(x_k))\right]$$

But we have:

$$f(x_k) - f(x_k - \nabla_i f(x_k)) \geq \frac{1}{2L_i} \|\nabla_i f(x_k)\|^2$$

where the last result follows from the fact that:

$$f(x + U_i h_i) \leq f(x) + \langle\, f_i^{'}(x), h_i \rangle + \frac{L_i}{2} \|h_i\|$$

which can be obtained by a Taylor expansion with integral remainder. Hence :

$$f(x_k) - \mathbb{E}_{i_k}(f(x_{k+1})) \geq \sum_{i=1}^n \frac{p_\alpha^{(i)}}{2L_i} \|f_i{}'(x)\|^2 = \frac{1}{2S_\alpha} \|f^{'}(x_k)\|_{1-\alpha}$$

Let X^* be the optimality set (i.e. the set of all the minimizers of f). Then we have, by convexity of f, and noting that $\forall x_* \in X^*, f(x_*) = f(x^*)$:

$$f(x_k) - f(x_k) = f(x_k) - f(x_*) \leq \min_{x_* \in X^*} \langle\, \nabla f(x_k), x_k - x_* \rangle$$
$$\leq \|\nabla f(x_k)\|_{1-\alpha} R_{1-\alpha}(x_0)$$

where the last inequality follows from the fact that:

- $\langle x, y \rangle \leq \|x\|_{1-\alpha} \|y\|_{1-\alpha}^*$
- $f(x_k) \leq f(x_0)$

Noting $C = 2S_\alpha R_{1-\alpha}^2(x_0)$, we then have:

$$f(x_k) - \mathbb{E}_{i_k}(f(x_{k+1})) \geq \frac{1}{C}(f(x_k) - f(x^*))^2$$

Then, taking the expectation in ξ_{k-1} and noting $\phi_k = \mathbb{E}_{\xi_{k-1}}(f(x_k))$, we have :

$$\phi_k - \phi_{k+1} \geq \frac{1}{C}(\phi_k - f(x^*))^2$$

Then, by some clever manipulation and using a telescopic sum of the term $\frac{1}{\phi_k - f(x^*)}$, we find that:

$$\phi_k - f(x^*) \leq \frac{C}{k+4} = \frac{2S_\alpha R_{1-\alpha}^2(x_0)}{k+4}$$

which is the desired result. ■

To compare this result with the full gradient method seen in class, we will examine two special cases of α.

- $\underline{\alpha = 0}$

This choice of α yields a remarkable algorithm where the choice of the index i_k is made uniformly at random. From the previous theorem, since $S_0 = n$, the RCDM($0, x_0$) bound writes:

$$\mathbb{E}_{\xi_{k-1}}(f(x_k) - f(x^*)) \leq \frac{2n}{k+4} R_{1-\alpha}^2(x_0)$$

Let's compare this result with the FGM bound. First, we give elements of proof of this bound. To that end, suppose that f is twice differentiable.

Suppose we have γ such that:

$$f^{''}(x) \preceq \gamma \text{ diag}\{L_i I_{n_i}\}_{i=1}^n$$

Then, we can show that this is equivalent to :$\forall y \in \mathbb{R}^N, y^T f^{''}(x) y \leq \gamma \|y\|_1^2$ (where the norm is the $\|.\|_\alpha$ norm defined earlier). This in turn implies that

$$\|\nabla f(x) - \nabla f(y)\|_{[1]}^* \leq \gamma \|x - y\|_{[1]}^2$$
$$\text{and then } f(y) \leq f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{1}{2}\gamma \|x - y\|_{[1]}^2$$

Then, following a rather similar proof than that of RCDM(α, x_0), we get the following bound for the FGM:

$$f(x_k) - f(x^*) \leq \frac{\gamma}{k} R_1^2(x_0)$$

Note that L_i is an upper bound of $f_i^{''}(x)$, i.e: $f^{''}(x)_i \preceq L_i I_{n_i}$. Then we can show that (Lemma 1, Nesterov, 2012) $f^{''}(x) \preceq n$ diag $(f^{''}(x_i)) \preceq n$ diag $\{L_i I_{n_i}\}$. Hence, in the worst case, we have $\gamma = n$ and then, in the worst case, the convergence rates of RCDM and FGM are proportional, which cumforts the use of RCDM since one iteration of RCDM is likely to be much cheaper than one of FGM.

- $\underline{\alpha = 1}$

Then for RCDM(1, x_0), we have:

$$\mathbb{E}_{\xi_{k-1}}(f(x_k) - f(x^*)) \leq \frac{2n}{k+4} \left[\frac{1}{n} \sum_{i=1}^n L_i\right] R_0^2(x_0)$$

Suppose now that a γ satisfies : $f^{''}(x) \preceq \gamma I_N$. Then we have :

$$f(x_k) - f(x^*) \leq \frac{\gamma}{k} R_0^2(x_0)$$

The maximal eigenvalue of $f^{''}(x)$ satisfies the latter condition. Since this value can reach the trace of the symmetric matrix $f^{''}(x)$, and that $\sum_{i=1}^n L_i$ is an upper bound on the diagonal elements of the Hessian, we can have in the worst case : $\gamma = \sum_{i=1}^n L_i$. Hence, in the worst case, the rates of convergence of RCDM and FGM are the same. Once again, since RCDM's iterations can be much cheaper, RCDM is preferable in the worst case.

These comparisons seem promising. But there remains some uncertainty in the fact that the convergence results of RCDM are in expectation. To show that the bounds are tight enough, Nesterov derives results in high probability through strongly convex functions' minimization.

C. Results in high probability using a strongly convex regularizer

In this section, we define the strong convexity with respect to the $\|.\|_{[1-\alpha]}$ norm : f is strongly convex with strong convexity parameter σ w.r.t the $\|.\|_{[1-\alpha]}$ norm if:

$$\forall x, y \in \mathbb{R}^N, f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2}\|y - x\|_{[1-\alpha]}$$

Minimizing both sides in y, we can show that:

$$\forall x \in \mathbb{R}^N, f(x) - f(x^*) \leq \frac{1}{2\sigma}(\|\nabla f(x)\|^*_{[1-\alpha]})^2$$

A simple application of this inequality with $\{x_k\}$ resulting from RCDM(α, x_0), after substracting $f(x^*)$ from both sides of the inequality and rearranging the inequality, it follows by taking the expectation w.r.t ξ_{k-1} in both sides that:

$$\mathbb{E}_{\xi_{k-1}}(f(x_k) - f(x^*)) \leq (1 - \frac{\sigma}{S_\alpha})^k (f(x_0) - f(x^*))$$

Let $\mu > 0$. We will now apply the previous strong convexity results on the regularized function : $f_\mu(x) = f(x) + \frac{\mu}{2}\|x - x_0\|^2_{[\alpha]}$. By applying RCDM(0, x_0) on f_μ, we can show that with great probability, we can have a very tight error bound on when minimizing f.

To that end, the central lemma we use is the following:

***Proposition** 2:* Let $\{x_k\}$ be the sequence generated by RCDM(0, x_0) applied to f_μ. Then:

$$\mathbb{E}_{\xi_k}(\|\nabla f_\mu(x_k)\|^*_{[1]}) \leq \left[2(n + \mu)(f(x_0) - f(x^*))(1 - \frac{\mu}{n})^k\right]^{1/2}$$

Let's now fix the desired accuracy of the solution $\epsilon > 0$ and confidence level $\beta \in (0, 1)$. The following theorem shows that the dependence of the complexity bounds in the confidence level β is very moderate.

***Theorem** 2:* Let $\mu = \frac{\epsilon}{4R_1^2(x_0)}$ and choose

$$k \geq 1 + \frac{8nR_1^2(x_0)}{\epsilon} \log \frac{2nR_1^2(x_0)}{\epsilon(1-\beta)}$$

Then :

$$\mathbb{P}(f(x_k) - f(x^*) \leq \epsilon) \geq \beta$$

The quantity $O(\frac{nR_1^2(x_0)}{\epsilon} \log \frac{nR_1^2(x_0)}{\epsilon(1-\beta)})$ of RCDM is to be compared with that of FGM which is $O(\frac{nR_1^2(x_0)}{\epsilon})$. Hence, up to a logarithmic factor, the complexity bounds of RCDM and FGM coincide with high probability.

D. Other results

In his article, Y. Nesterov further develops bounds on constrained minimization problems. He also extends usual accelerated gradient methods to randomized coordinate descent, and shows that the acceleration indeed leads to tighter bounds, but that sometimes this acceleration can't be implemented efficiently, which can make RCDM useless compared to full gradient methods. Also, throughout the paper, it is assumed that tight Lipschitz constants are known, which is not always the case in practice. He develops a method for dynamically adjusting these constants, but warns that the computational complexity of this method can be prohibitive.

III. Experiments

To show the benefits of the method, we have chosen to implement RCDM$(1, 0)$ on a simple function : $f(x) = \frac{1}{2}\|Ax - b\|^2$ where

- the known solution $x^* \in \mathbb{R}^{1000}$ is chosen randomly from a $\mathcal{N}(0, 1)$
- $b = Ax^*$
- the matrix $A \in \mathbb{R}^{300\times1000}$ is samples of a centered Gaussian vector with covariance given by the Toeplitz matrix, which allows to test the performance of the algorithms for easy and difficult cases of the function by choosing different correlation values ρ of the entries of A. Furthermore, the rows of A are equally sparse, with a third of the entries of each row of A being non-zero. The choice of such a matrix is motivated by the fact that we can implement efficiently coordinate descent methods by smart updates at each iteration. Indeed, the gradient of f writes : $\nabla f(x) = A^T(Ax - b)$, while the coordinate-wise gradient is : $e_i^T A^T(Ax - b)$. Computing it amounts to: (i) updating the residuals $r = Ax - b$, which can be reduced to one scalar-vector product and a sum since we have: $r_{k+1} = r_k + (x_{k+1}^{i_k+1} - x_k^{i_k+1})A_{i_k,:}$ (ii) computing one vector-vector product : $A_{i_k,:}r_k$

Since time plotting would be too dependent upon the efficiency of the implementation of both RCDM and FGM, and to guarantee the impartiality of the results, we compare the two algorithms in terms of number of full iterations: for coordinate descent, one full iteration corresponds to one full pass (1000 iterations) over the data (note that some coordinates may not be visited since we draw from a probability distribution).We verify on Figure 1 that the convergence of RCDM in terms of full iterations is much faster than that of FGM.

To comfort the results found in Section II.A, we compare the effect of changing the correlation of the columns of A on RCDM and FGM. This is motivated by the fact that $f''(x) = A^TA$ and, noting $\lambda_{\max}$ the highest eigenvalue of A^TA, we have : $L = \lambda_{\max}$ and $L_i = \|A_{:,i}\|^2$. Then, the higher the correlation between A s columns, the higher $\lambda_{\max}$, and the closer it is to the Hessian's trace. We then expect FGM to be much more affected by the increase in the correlation than RCDM. This is verified on Figure 2.

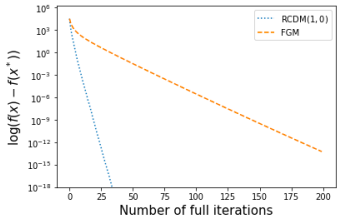

Figure 1: Comparison of RCDM(1, 0) and FGM. $\rho = 0.5$

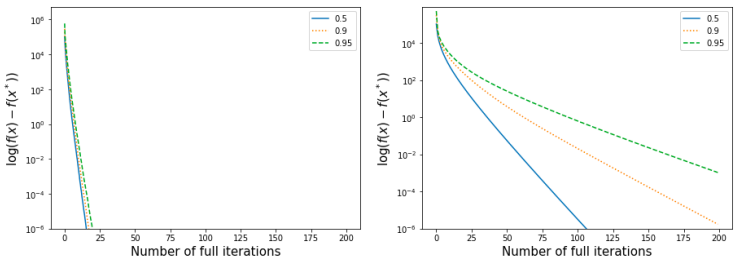

Figure 2 : Comparison of RCDM(1, 0) and FGM for different values of ρ. Left: RCDM(1, 0). Right: FGM.

References

[1] A. Beck and L. Tetruashvili, On the convergence of block coordinate descent type methods. SIAM J. OPTIM, Vol. 23, No. 4, pp. 20372060.

[2] Y. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim., 22(2), 341362