COORDINATE DESCENT ALGORITHMS

ZINEB SORDO, ERIC CHAGNON, AND RICHARD LY

ABSTRACT. Coordinate Descent algorithms may be used to solve optimization problems by iteratively moving along coordinate directions or coordinate hyperplanes that approximately minimize the objective function. The algorithm and its number of variants have numerous applications, enabling it to see a continued growth in popularity in topics such as machine learning and data analysis, among others. In this paper, we introduce two such Coordinate Descent variants, as well as their convergence rates under certain assumptions.

1. Introduction

Coordinate Descent algorithms are an iterative approach to solving optimization problems. In each iteration, a number of the components in the variable vector x are held constant, and the objective function is approximately minimized with respect to the remaining components. In doing so, this creates subproblems which are lower dimensional (or potentially scalar) minimization problems. These subproblems are usually more easily solved than the original, full optimization problem.

The Coordinate Descent algorithms we discuss here are concerned with the unconstrained minimization problem:

$$\min_x f(x), \text{ where } f : \mathbb{R}^n \to \mathbb{R} \text{ is continuous} \tag{1}$$

The assumptions made on f differ depending on the variant of CD algorithm. Variants of the CD algorithm are defined by the way in which the index $i_k \in \{1, 2, \dots, n\}$ is selected in each iteration k. The index i_k is used to select the coordinate of the gradient $\nabla f(x)$ used to update x^k and obtain x^{k+1}. In addition to describing the algorithms in Section 2, we will also address the convergence rate of these variants towards the minimizer x^* of the objective function $f(x)$ in Section 3. We give key proof ideas for their respective convergence rates in Section 4, and detail their full proofs in Section 5.

We consider two fundamental Coordinate Descent algorithm variants: Randomized and Cyclic. These variants are concerned with adjusting just one coordinate in each iteration, as opposed to block variants in which groups of blocks of indices are changed in each iteration. In other words, the Randomized and Cyclic variants we later discuss search along a single coordinate direction rather than a coordinate hyperplane in their approach towards the minimizer x^*.

Although there is a defining difference between the Randomized and Cyclic variants that lies in the selection process of which coordinate direction to optimize along, we will see that the general approach taken in each iteration of their respective algorithms are similar.

2. Algorithm Description

Randomized CD algorithm

Algorithm 1 Randomized Coordinate Descent

```
choose $x^0 \in \mathbb{R}^n$;
set $k \leftarrow 0$;
repeat
    Choose index $i_k$ with uniform probability from $\{1, 2, \dots, n\}$ independent of previous iterations;
    Set $x^{k+1} \leftarrow x^k - \alpha_k [\nabla f(x)]_{i_k} e_{i_k}$ for some $\alpha_k > 0$;
    $k \leftarrow k + 1$;
until termination test satisfied
```

The Randomized CD algorithm begins with an arbitrary choice $x^0 \in \mathbb{R}^n$. For each iteration k, an index $i_k \in \{1, 2, \dots, n\}$ is chosen with uniform probability. This choice of i_k does not depend on the selections of i_k in previous iterations. Having selected an index i_k, the i_k^{th} coordinate of $\nabla f(x)$, denoted by $[\nabla f(x)]_{i_k}$, is the direction moved along with step size $\alpha_k > 0$ to update x^k and obtain x^{k+1}. The step size α_k may be chosen to obtain varying convergence results. This process is repeated until a termination test is satisfied, such as achieving a desired level of convergence or reaching a maximum number of iterations.

Cyclic CD algorithm

Algorithm 2 Cyclic Coordinate Descent

```
choose $x^0 \in \mathbb{R}^n$;
set $k \leftarrow 0$;
set $i_0 \leftarrow 1$;
repeat
    Set $x^{k+1} \leftarrow x^k - \alpha_k [\nabla f(x)]_{i_k} e_{i_k}$ for some $\alpha_k > 0$;
    $i_{k+1} \leftarrow (i_k \bmod n) + 1$;
    $k \leftarrow k + 1$;
until termination test satisfied
```

Alternatively, in the cyclic variant of Algorithm 1 of [1], the component i_k of the gradient $\nabla f(x)$ is selected in a cyclic fashion, in which $i_0 = 1$ and $i_{k+1} = [i_k \bmod n] + 1, k = 0, 1, 2....$ And each selected components is modified at least once for every T iterations for some $T \geq n$, such that:

$$\cup_{j=0}^{T} \{i_{k-j}\} = 1, 2, 3, ..., n \text{ for all } k \geq T$$

Thus, the cyclic CD algorithm is similar to Algorithm 1 except for i_k being selected in a cyclic fashion, and $\alpha_k \equiv \frac{1}{L_{max}}$ with L_{max} being the maximal absolute value of the diagonal elements of the Hessian, i.e. $L_max = \max_{i=1,2,\dots,n} L_i$ where L_i are the component Lipschitz constants and L is the standard Lipschitz constant.

3. Main Results

Randomized CD algorithm

The following assumptions are required to ensure the convergence of the Randomzed CD algorithm:

- The function is convex,
- The function is uniformly Lipschitz continuously differentiable
- The function attains its minimum value f^* on a set $\boldsymbol{S}$
- $\alpha_k \equiv \frac{1}{L_{max}}$ in Algorithm 1 with i_k selected in an iid. uniform way with replacement

When these assumptions are satisfied the following is true:

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

In the case of strong convexity:

$$E[f(x^k)] - f^* \leq ((1 - \frac{\sigma}{nL_{\max}}))^k((f(x^0) - f^*) \tag{3}$$

Cyclic CD algorithm

Based on [Wri15] and [Bec13], the assumption that should be met to ensure convergence are similar as the ones used for the Randomized CD algorithm except for the 4th assumption which is in rather:

- $\alpha_k \equiv \frac{1}{L_{max}}$ in Algorithm 1 with i_k selected in the cyclic fashion previously described $\nabla^2 f(x)$

Then assuming these assumptions are met, for $k = n, 2n, 3n, ...,$, we have:

$$f(x^k) - f^* \leq \frac{4nL_{max}(1 + n\frac{L^2}{L^2_{max}})R_0^2}{k+8} \tag{4}$$

And in the case of strong convexity, then:

$$f(x^k) - f^* \leq (1 - \frac{\sigma}{2L_{max}(1 + n\frac{L^2}{L^2_{max}})})^{k/n}(f(x^0) - f^*) \tag{5}$$

Convergence properties comparisons

In the case of nonconvex functions, convergence is not always guaranteed as seen in Powell's Example in [Pow73] where small perturbations lead to a nonconvergence behavior in the Cyclic method as well as in the Randomized approach. According to [Ber99][4, Proposition 2.7.1] Cyclic convergence for nonconvex functions is possible only if the minimizer along any coordinate direction from any point is unique.

In terms of algorithm complexity, we can see that in (4), since $L \geq L_{max}$ the numerator is $\mathcal{O}(n^2)$ compared to the Randomized Algorithm in (2) where the complexity is $\mathcal{O}(n)$. The same difference of n is seen in the case of strongly convex function if we make the approximation of $(1-\epsilon)^{1/n} \approx 1 - \epsilon/n$ for ϵ close to 0 as n increases and with $\epsilon = \frac{\sigma}{2L_{max}(1+n\frac{L^2}{L^2_{max}})}$.

In addition, taking larger values of L_{max}, as long as they are defined as the coordinate Lipschitz positive constant, will lead to shorter iteration steps as $\alpha_k = 1/L_{max}$ and therefore to a different larger iteration complexity.

After simulations were made for both variants on a quadratic function $f(x) = \frac{1}{2}x^TQx$ with Q symmetric positive semidefinite with its maximum diagonal being L_{max}, and with a fixed steplength ($\alpha_k \equiv 1/L_{max}$) and an optimal steplength $\alpha_k \equiv 1/Q_{i_k,i_k}$ the following was highlighted:

- If initialization and conditions are not properly set up, $f(x^k)$ shows a rapid followed by a linear rate decrease.
- In the case of Randomized CD, the version including sampling without replacement converges faster than the version with sampling with replacement.
- The optimal steplength could enable to up to 6 times fewer iterations the fixed steplength
- The Cyclic variant is more sensitive to small changes in the settings which makes is a lot slower than the randomized variant by factors of at least 10.

4. Key Proof Ideas

Assumption 1 Given the function in (1) is convex and uniformly Lipschitz continuously differentiable, and attains its minimum value f^* on a set S. There is a finite R_0 such that the level set for f defined by x^0 is bounded that is:

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

Using **Assumption 1** and Lemmas listed in the appendix we can easily see that

$$f(x^{k+1}) = f(x^k - \alpha_k [\nabla f(x^k)]_{i_k} e_{i_k})$$

can be re-written as

$$E_{i_k}[f(x^{k+1})] \leq f(x^k) - \frac{1}{2nL_{\max}} \|\nabla f(x^k)\|^2$$

Subtract $f(x^*)$ from both sides and let $\phi_k = E[f(x^k)] - f(x^*)$ and take expectation of both sides with respect to all variables

$$\phi_{k+1} \leq \phi_k - \frac{1}{2nL_{\max}} E[\|\nabla f(x^k)\|^2]$$

By Jensen's Inequality

$$\phi_{k+1} \leq \phi_k - \frac{1}{2nL_{\max}} [E[\|\nabla f(x^k)\|]]^2$$

Using the result from **Lemma 4** and some re-writing of the above equation yields the following convergence result:

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

When f is strongly convex a similar approach is used with a different initial step that involves taking the minimum of both sides of **Lemma 3** with respect to y and setting $x = x^k$. This gives the final convergence result for strongly convex f:

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

Cyclic variant

Assuming **Assumption 1** and with i_k chosen in a cyclic fashion, we use the following theorem 3.6 of [Bec13] (Cyclic CD by block):

$$f(x^k) - f^* \leq \frac{4\overline{L}_{max}(1 + p\frac{L^2}{\overline{L}^2_{min}})R_0^2}{k + 8/p} \tag{6}$$

Then, by updating $\overline{L}_{min} = \overline{L}_{max} = L_{max}$, $k = n, 2n, ...,$ and $p = n$, we have result (4).

A similar approach is taken for result (5) by applying theorem 3.9 of [Bec13] for the strongly convex case with the same updates and with a parameters $\sigma > 0$

5. Full Proof

Some additional facts used in the proof are provided in the appendix.

Given **Assumption 1**

First we select stepsize of $\alpha_k = 1/L_{\max}$, thus have the following:

$$f(x^{k+1}) = f(x^k - \alpha_k[\nabla f(x^k)]_{i_k} e_{i_k})$$

Randomized CD

Using Taylor's Theorem, **Lemma 1** and **Lemma 2**:

$$f(x^{k+1}) \leq f(x^k) - \alpha_k[\nabla f(x^k)]^2_{i_k} + \frac{1}{2}\alpha_k^2 L_{i_k}[\nabla f(x^k)]^2_{i_k}$$

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

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

Take the expectation of both sides over the randomly selected index i_k.

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

$$E_{i_k}[f(x^{k+1})] \leq f(x^k) - \frac{1}{2nL_{\max}}\|\nabla f(x^k)\|^2$$

Subtract $f(x^*)$ from both sides and let $\phi_k = E[f(x^k)] - f(x^*)$ and take expectation of both sides with respect to all variables

$$\phi_{k+1} \leq \phi_k - \frac{1}{2nL_{\max}}E[\|\nabla f(x^k)\|^2]$$

By Jensen's Inequality

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

Using the result from **Lemma 3** we can plug that into (6) and get the following:

$$\phi_k - \phi_{k+1} \geq \frac{1}{2nL_{\max}}\phi_k^2\frac{1}{R_0^2}$$

$$\frac{\phi_k - \phi_{k+1}}{\phi_k^2} \geq \frac{1}{2nL_{\max}}\frac{1}{R_0^2}$$

Since $\phi_k > \phi_{k+1}$

$$\frac{\phi_k - \phi_{k+1}}{\phi_k\phi_{k+1}} \geq \frac{\phi_k - \phi_{k+1}}{\phi_k^2} \geq \frac{1}{2nL_{\max}}\frac{1}{R_0^2}$$

$$\frac{1}{\phi_k} - \frac{1}{\phi_{k+1}} = \frac{\phi_k - \phi_{k+1}}{\phi_k\phi_{k+1}}$$

Therefore,

$$\frac{1}{\phi_{k+1}} - \frac{1}{\phi_k} \geq \frac{1}{2nL_{\max}R_0^2}$$

By applying this recursively we obtain:

$$\frac{1}{\phi_k} \geq \frac{1}{\phi_0} + \frac{1}{2nL_{\max}R_0^2} \geq \frac{k}{2nL_{\max}R_0^2}$$

Inverting both sides and taking the expectation yields:

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

For strong convexity we follow a similar procedure but begin by taking the minimum of both sides of **Lemma 4** with respect to y and setting $x = x^k$

$$f^* \geq f(x^k) - \frac{1}{2\sigma}\|f(x^k)\|^2$$

Use the result from **Lemma 3** and plug into the above inequality:

$$\phi_{k+1} \leq \phi_k - \frac{\sigma}{nL_{\max}}\phi_k = (1 - \frac{\sigma}{nL_{\max}})\phi_k$$

By applying this function recursively we obtain:

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

Cyclic CD

Using the **Descent lemma**, we have that:

$$f(x^0) - f^* \leq \frac{L}{2}||x^0 - x^*||^2 \leq \frac{L}{2}R_0^2$$

Thus,

$$f(x^0) - f^* \leq \frac{L}{2}R_0^2$$

And with $L \leq nL_{max}$, we have:

$$\frac{L}{2}R_0^2 \leq \frac{nL_{max}}{2} \leq \frac{n}{8}(4L_{max}(1 + nL^2/L_{max}^2)R_0^2) = \frac{4nL_{max}}{8}(1 + \frac{nL^2}{L_{max}^2})R_0^2$$

Then we have for a generated sequence $\{x^k\}_{k\geq 0}$ by the BCGD method of [Bec13]:

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{4L_{max}(1 + \frac{nL^2}{L_{max}^2})R_0^2}(f(x^k) - f^*)^2 \geq \frac{1}{4L_{max}(1 + \frac{nL^2}{L_{max}^2})R_0^2}(f(x^k) - f^*)^2$$

The next step uses the following Lemma:

For a nonnegative sequence of real numbers $\{A_k\}_{k\geq 0}$ satisfying $A_k - A_{k+1} \geq \gamma A_k^2$, k = 0,1,... and $A_0 \leq \frac{1}{m\gamma}$ then for some positive γ and m, we have:

$$A_k \leq \frac{1}{\gamma}.\frac{1}{k+m}$$

and appyling it with $\gamma = \frac{1}{4(L_{max}+\frac{nL^2}{L_{max}})R_0^2}$, $m = 8/n$ and $k = n, 2n, ...$, leads to the final result (4).

In the case, of strong convexity of f with parameter $\sigma > 0$, by minimizing both side of the following, for every x,y real numbers, and with respect to y:

$$f(y) \geq f(x) + < \nabla f(x), y - x > + \frac{\sigma}{2}||y-x||^2$$

We obtain for every real number x:

$$f(x) - f* \leq \frac{1}{2\sigma}||\nabla f(x)||^2 \tag{8}$$

Then, combining the previous result with: $f(x_k) - f(x_{k+1}) \geq \frac{1}{4L_{max}(1+nL^2/L_{max}^2)}||\nabla f(x_k)||^2$ we finally obtain results (5).

References

[Bec13] Tetruashvili L. Beck, A., *On the convergence of block coordinate descent methods.*, SIAM Journal on Optimization **23** (2013), no. 4, 2037–2060.

[Ber99] D.P. Bertsekas, *Nonlinear programming, second edn.*

[Pow73] M.J.D Powell, *On search directions for minimization algorithms*, Mathematical Programming **4** (1973), 193–201.

[Wri15] Stephen J. Wright, *Coordinate descent algorithms*, Proceedings of the National Academy of Sciences (2015).

Appendix A. Additional Details of the Proofs

Lemma 1

$$\|\nabla[f(x+te_i)]_i - [\nabla f(x)]_i\| \leq L_i\|t\| \tag{9}$$

Lemma 2

$$L_{\max} = \max_{i=1,2,..,n} L_i \tag{10}$$

Lemma 3
When f is strongly convex:

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\sigma}{2}\|y-x\|_2^2 \tag{11}$$

Lemma 4
Given that f is convex, for any $x^* \in S$, then using **Assumption 1**:

$$f(x^k) - f^* \leq \nabla f(x^k)^T(x^k - x^*) \leq \|\nabla f(x^k)\|\|x^k - x^*\| \leq R_0\|\nabla f(x^k)\|$$

$$(f(x^k) - f^*)\frac{1}{R_0} \leq \|\nabla f(x^k)\|$$

Taking the expectation of both sides yields:

$$E[\|\nabla f(x^k)\|] \geq \phi_k \frac{1}{R_0} \tag{12}$$

Descent Lemma
Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a continuous differentiable function whose gradient ∇g is Lipschitz with constant M. Then,

$$g(y) \leq g(x) + < \nabla g(x), y-x > + \frac{M}{2}x - y^2 \; for\ all\ x, y \in \mathbb{R}^d \tag{13}$$

Email address: zsordo@ucdavis.edu

Email address: echagnon@ucdavis.edu

Email address: rkly@ucdavis.edu