

Coordinate Descent Methods for Huge-Scale Optimization

Martin Takáč

First-Year Report
Graduate School of Mathematics
University of Edinburgh
August 2011

Abstract

In the first chapter of this report I briefly describe my participation at doctoral and other courses, seminars and workshops, my presentations at seminars and conferences, summarize my research output, outline the areas of my current research, and conclude with a list of awards and service. The final two parts contain the abstracts of my papers [17, 16]; in the case of the former I include an excerpt from the paper as well.

Contents

Abstract	2
1 Overview of My First Year of PhD	4
1.1 Attendance at Workshops, Courses and Seminars	4
1.2 Talks at Seminars	5
1.3 Talks at Conferences	5
1.4 Current Research	6
1.5 Service	6
1.6 Awards	6
2 Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function (Abstract and Excerpt)	7
2.1 Introduction	8
2.2 Problem Description and Our Contribution	9
2.3 Assumptions and the Algorithm	12
2.4 Coordinate Descent for Composite Functions	15
2.4.1 Convex Objective	16
2.4.2 Strongly Convex Objective	18
3 Efficient Serial and Parallel Coordinate Descent Methods for Huge-Scale Truss Topology Design (Abstract)	20

Chapter 1

Overview of My First Year of PhD

During the first year of my PhD I dedicated about 20% of my working time on literature review of Optimization techniques, concretely I focused on Coordinate Descent Methods (CD). I have read more than 30 papers. The key papers in my research so far were [13, 12, 33, 24, 8, 25].

1.1 Attendance at Workshops, Courses and Seminars

As a requirement for all first year mathematics PhD students in Scotland I took Scottish Mathematical Sciences Training Centre (SMSTC) courses. I took the Probability and Applied Mathematics Methods streams. I am happy to report that I have got A grades in both streams. I also did a Reading Course from book [5] organized by Dr. Richtarik where I have been working through all exercises (100+ of them) and I have got A grade.

I also attended multiple workshops and courses:

- *2011, September 12–16, Combinatorial Optimization, NATCOR Course, University of Southampton* (upcoming course). NATCOR Courses are similar to SMSTC, but the topics come from optimization areas.
- *2011, May 12–13, Computational Complexity Challenges in Optimization*, Edinburgh University (Keynote speaker: Felipe Cucker).
- *2011, May 6, Proof Reading (Generic Skills Courses)*, Linguistics and English Language, Edinburgh.
- *2011, April 4–8, Stochastic Modeling*, NATCOR Course, Lancaster University.
- *2011, February 17–18, Engaging with Engagement (Generic Skills for PG students)*, 15 South College Street, Edinburgh.
- *2011, February 16, Generic Skills Course for Mathematics Postgraduates*, SMSTC, Dundee.

- 2011, January 19, **EUSA Inspiring Teaching Conference**, Edinburgh.
- 2010, December 10, **Fragility and Robustness of Networks**, Edinburgh.
- 2010, November 8–10, **Parallel Programming with MPI (by NAG)**, Edinburgh.
- 2010, November 11–12, **OpenMP (by NAG)**. This two days workshop was a nice introduction into parallel programming using OpenMP in C. I have found this workshop really useful.
- 2010, October 26, **nVIDIA GPU Workshop**. On this workshop, there were explained the GPU architecture, hierarchy of memories and CUDA.

I have also attended multiple talks organized by the Edinburgh Research Group in Optimization (ERGO).

1.2 Talks at Seminars

I regularly attended the Sparsity Reading Group and I also presented there Nesterov's paper *Efficiency of coordinate descent methods on huge-scale optimization problems* [13]. I have also delivered talks at 2 seminars; one more is upcoming:

- 2011, October, **OR Society talk** (upcoming).
- 2011, June, 23, **PhD colloquium**.
- 2011, July, 28, **EUSci seminar**, This was the most challenging one, because I have presented it for non-mathematical audience and therefore it had to be prepared more carefully .

1.3 Talks at Conferences

I have attended multiple conferences, where I gave a talk or presented a poster:

- 2011, September 28–30, **Facing the Mulicore-Challenge II**, Conference for Young Scientists, September 28-30, 2011 Karlsruhe Institute of Technology (KIT), Germany (talk) (upcoming).
- 2011, August 30–September 2, **International Conference On Operations Research**, Zurich, Switzerland (talk) (Invited by Yurii Nesterov) (upcoming).
- 2011, July 1, **24th Biennial Conference on Numerical Analysis**, Strathclyde, Glasgow (talk) (Invited by Dr. Richtarik).
- 2011, June 27–30, **Workshop : Signal Processing with Adaptive Sparse Structured Representations**, Edinburgh (talk).

- *2011, May 16–19, SIAM Conference on Optimization*, Darmstadt, Germany (poster).
- *2011, April 11–13, LANCS Workshop on Modeling and Solving Complex Optimisation Problems*, Lancaster (talk).
- *2011, February 01, Edinburgh SIAM Student Chapter Conference*, Edinburgh (talk and poster).

1.4 Current Research

With my supervisor Dr. Peter Richtárik we have written following papers (technical reports):

- *P. Richtárik and M. Takáč*: Efficient serial and parallel coordinate descent methods for huge-scale truss topology design [16].
- *P. Richtárik and M. Takáč*: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function [17].
- *P. Richtárik and M. Takáč*: Efficiency of randomized coordinate descent methods on minimization problems with a composite objective function [15].

1.5 Service

I was also involved with High School teaching organized by School of Mathematics. Concretely, I helped on these following events:

- *2011, June 18*, Stirling masterclass.
- *2011, June*, Tutoring on LEAPS summer school mathematics course. I have delivered multiple tutorials.
- *2011, May*, Higher Mathematics Revision.

1.6 Awards

- *2011, July, 01*, Certificate of Appreciation, 24th Biennial Conference on Numerical Analysis, Strathclyde, Glasgow.
- *2011, February 01*, Best poster award, Edinburgh SIAM Student Chapter Conference, Edinburgh.

Chapter 2

Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function (Abstract and Excerpt)

Abstract In this paper we develop a randomized block-coordinate descent method for minimizing the sum of a smooth and a simple nonsmooth block-separable convex function and prove that it obtains an ϵ -accurate solution with probability at least $1 - \rho$ in at most $O(\frac{n}{\epsilon} \log \frac{1}{\rho})$ iterations, where n is the number of blocks. For strongly convex functions the method converges linearly. This extends recent results of Nesterov [Efficiency of coordinate descent methods on huge-scale optimization problems, CORE Discussion Paper #2010/2], which cover the smooth case, to composite minimization, while at the same time improving the complexity by the factor of 4 and removing ϵ from the logarithmic term.

More importantly, in contrast with the aforementioned work in which the author achieves the results by applying the method to a regularized version of the objective function with an unknown scaling factor, we show that this is not necessary, thus achieving true iteration complexity bounds. In the smooth case we also allow for arbitrary probability vectors and non-Euclidean norms.

Finally, we demonstrate numerically that the algorithm is able to solve huge-scale ℓ_1 -regularized least squares and support vector machine problems with a billion variables.

2.1 Introduction

The basic algorithmic strategy of CD methods is known in the literature under various names such as alternating minimization, coordinate relaxation, linear and non-linear Gauss-Seidel methods, subspace correction and domain decomposition. As working with all the variables of an optimization problem at each iteration may be inconvenient, difficult or impossible for any or all of the reasons mentioned above, the variables are partitioned into manageable blocks, with each iteration focused on updating a single block only, the remaining blocks being fixed. Both for their conceptual and algorithmic simplicity, CD methods were among the first optimization approaches proposed and studied in the literature (see [1] and the references therein; for a survey of block CD methods in semidefinite programming we refer the reader to [26]). While they seem to have never belonged to the mainstream focus of the optimization community, a renewed interest in CD methods was sparked recently by their successful application in several areas—training support vector machines in machine learning [6, 3, 19, 31, 30], optimization [9, 24, 23, 22, 33, 18, 13, 27], compressed sensing [8], regression [29], protein loop closure [2] and truss topology design [16]—partly due to a change in the *size* and *nature of data* described above.

Order of coordinates. Efficiency of a CD method will necessarily depend on the balance between time spent on choosing the block to be updated in the current iteration and the quality of this choice in terms of function value decrease. One extreme possibility is a *greedy* strategy in which the block with the largest descent or guaranteed descent is chosen. In our setup such a strategy is prohibitive as i) it would require all data to be available and ii) the work involved would be excessive due to the size of the problem. Even if one is able to compute all partial derivatives, it seems better to then take a full gradient step instead of a coordinate one, and avoid throwing almost all of the computed information away. On the other end of the spectrum are two very cheap strategies for choosing the incumbent coordinate: *cyclic* and *random*. Surprisingly, it appears that complexity analysis of a cyclic CD method in satisfying generality has not yet been done. The only attempt known to us is the work of Saha and Tewari [18]; the authors consider the case of minimizing a smooth convex function and proceed by establishing a sequence of comparison theorems between the iterates of their method and the iterates of a simple gradient method. Their result requires an isotonicity assumption. Note that a cyclic strategy assumes that the data describing the next block is available when needed which may not always be realistic. The situation with a random strategy seems better; here are some of the reasons:

- (i) Recent efforts suggest that complexity results are perhaps more readily obtained for randomized methods and that randomization can actually improve the convergence rate [20, 7, 19].
- (ii) Choosing all blocks with equal probabilities should, intuitively, lead to similar results as is the case with a cyclic strategy. In fact, a randomized

strategy is able to avoid worst-case order of coordinates, and hence might be preferable.

- (iii) Randomized choice seems more suitable in cases when not all data is available at all times.
- (iv) One may study the possibility of choosing blocks with different probabilities. The goal of such a strategy may be either to improve the speed of the method, or a more realistic modeling of the availability frequencies of the data defining each block.

Step size. Once a coordinate (or a block of coordinates) is chosen to be updated in the current iteration, partial derivative can be used to drive the step length in the same way as it is done in the usual gradient methods. As it is sometimes the case that the computation of a partial derivative is *much cheaper and less memory demanding* than the computation of the entire gradient, CD methods seem to be promising candidates for problems described above. It is important that line search, if any is implemented, is very efficient. The entire data set is either huge or not available and hence it is not reasonable to use function values at any point in the algorithm, including the line search. Instead, cheap partial derivative and other information derived from the problem structure should be used to drive such a method.

2.2 Problem Description and Our Contribution

The problem. We study the *iteration complexity* of simple randomized block coordinate decent methods applied to the problem of minimizing a *composite objective function*, i.e., a function formed as the sum of a smooth convex and a simple nonsmooth convex term:

$$\min_{x \in \mathbf{R}^N} F(x) \stackrel{\text{def}}{=} f(x) + \Psi(x). \quad (2.1)$$

We assume that this problem has a minimum ($F^* > -\infty$), f has (block) coordinate Lipschitz gradient, and Ψ is a (block) separable proper closed convex extended real valued function (these properties will be defined precisely in Section 2.3). Possible choices of Ψ include:

- (i) $\Psi \equiv 0$. This covers the case of *smooth minimization*. Complexity results are given in [13].
- (ii) Ψ is the indicator function of a block-separable convex set (such as a box). This choice models *problems with constraints on blocks of variables*; iteration complexity results are given in [13].
- (iii) $\Psi(x) \equiv \lambda \|x\|_1$ for $\lambda > 0$. In this case we can decompose \mathbf{R}^N onto N blocks. Increasing λ encourages the solution of (2.1) to be sparser [28]. Applications abound in, for instance, machine learning [3], statistics [21] and signal processing [8].

- (iv) There are many more choices such as the elastic net [34], group lasso [32, 10, 14] and sparse group lasso [4].

Iteration complexity results. Strohmer and Vershynin [20] have recently proposed a randomized Kaczmarz method for solving overdetermined consistent systems of linear equations and proved that the method enjoys global linear convergence whose rate can be expressed in terms of the condition number of the underlying matrix. The authors claim that for certain problems their approach can be more efficient than the conjugate gradient method. Motivated by these results, Leventhal and Lewis [7] studied the problem of solving a system of linear equations and inequalities and in the process gave iteration complexity bounds for a randomized CD method applied to the problem of minimizing a convex quadratic function. In their method the probability of choice of each coordinate is proportional to the corresponding diagonal element of the underlying positive semidefinite matrix defining the objective function. These diagonal elements can be interpreted as Lipschitz constants of the derivative of a restriction of the quadratic objective onto one-dimensional lines parallel to the coordinate axes. In the general (as opposed to quadratic) case considered in this paper (2.1), these Lipschitz constants will play an important role as well. Lin et al. [3] derived iteration complexity results for several smooth objective functions appearing in machine learning. Shalev-Schwarz and Tewari [19] proposed a randomized coordinate descent method with uniform probabilities for minimizing ℓ_1 -regularized smooth convex problems. They first transform the problem into a box constrained smooth problem by doubling the dimension and then apply a coordinate gradient descent method in which each coordinate is chosen with equal probability. Nesterov [13] has recently analyzed randomized coordinate descent methods in the smooth unconstrained and box-constrained setting, in effect extending and improving upon some of the results in [7, 3, 19] in several ways.

While the *asymptotic convergence rates* of some variants of CD methods are well understood [9, 24, 23, 22, 33], *iteration complexity* results are very rare. To the best of our knowledge, randomized CD algorithms for minimizing a composite function have been proposed and analyzed (in the iteration complexity sense) in a few special cases only: a) the unconstrained convex quadratic case [7], b) the smooth unconstrained ($\Psi \equiv 0$) and the smooth block-constrained case (Ψ is the indicator function of a direct sum of boxes) [13] and c) the ℓ_1 -regularized case [19]. As the approach in [19] is to rewrite the problem into a smooth box-constrained format first, the results of [13] can be viewed as a (major) generalization and improvement of those in [19] (the results were obtained independently).

Contribution. We further improve upon and extend and simplify the iteration complexity results of Nesterov [13], treating the problem of minimizing the sum of a smooth convex and a simple nonsmooth convex block separable function (2.1). We focus exclusively on simple (as opposed to accelerated) methods. The reason for this is that the per-iteration work of the accelerated algorithm in [13] on huge scale instances of problems with *sparse* data (such as the Google problem where sparsity corresponds to each website linking only to a few other websites or the

sparse problems) is excessive. In fact, even the author does not recommend using the accelerated method for solving such problems; the simple methods seem to be more efficient.

Each algorithm of this work and in papers [16, 15, 17] is supported by a high probability iteration complexity result. That is, for any given *confidence level* $0 < \rho < 1$ and *error tolerance* $\epsilon > 0$, we give an explicit expression for the number of iterations k which guarantee that the method produces a random iterate x_k for which

$$\mathbf{Prob}(F(x_k) - F^* \leq \epsilon) \geq 1 - \rho.$$

Table 2.1 summarizes the main complexity results of this work. Algorithm 2—Uniform (block) Coordinate Descent for Composite functions (UCDC)—is a method where at each iteration the block of coordinates to be updated (out of a total of $n \leq N$ blocks) is chosen uniformly at random.

Algorithm	Objective	Complexity
Algorithm 2 (UCDC) (Theorem 2)	convex composite	$\frac{2n \max\{\mathcal{R}_L^2(x_0), F(x_0) - F^*\}}{\epsilon} (1 + \log \frac{1}{\rho})$ $\frac{2n \mathcal{R}_L^2(x_0)}{\epsilon} \log \left(\frac{F(x_0) - F^*}{\epsilon \rho} \right)$
Algorithm 2 (UCDC) (Theorem 4)	strongly convex composite	$\max\{\frac{4}{\mu}, \frac{\mu}{\mu-1}\} n \log \left(\frac{F(x_0) - F^*}{\rho \epsilon} \right)$

Table 2.1: Summary of complexity results obtained in this work.

The symbols $P, L, \mathcal{R}_W^2(x_0)$ and μ appearing in Table 2.1 will be defined precisely in further sections. For now it suffices to say that L encodes the (block) coordinate Lipschitz constants of the gradient of f , P encodes the probabilities $\{p_i\}$, $\mathcal{R}_W^2(x_0)$ is a measure of distance of the initial iterate x_0 from the set of minimizers of the problem (2.1) in a norm defined by W (see Section 2.3) and μ is the strong convexity parameter of F (see Section 2.4.2). In the nonsmooth case μ depends on L and the smooth case it depends both on L and P .

1. **Composite setting.** We consider the composite setting (2.1), whereas [13] covers the unconstrained and constrained smooth setting only.
2. **No need for regularization.** Nesterov’s high probability results in the case of minimizing a function which is not strongly convex are based on regularizing the objective to make it strongly convex and then running the method on the regularized function. Our contribution here is that we show that no regularization is needed by doing a more detailed analysis using a thresholding argument (Theorem 1).
3. **Better complexity.** Our complexity results are better by the constant factor of 4. Also, we have removed ϵ from under the logarithm.
4. **General probabilities.** Nesterov considers probabilities p_i proportional to L_i^α , where $\alpha \geq 0$ is a parameter. High probability results are proved in

[13] for $\alpha \in \{0, 1\}$ only. Our results in the smooth case hold for an arbitrary probability vector p .

5. **General norms.** Nesterov's expectation results (Theorems 1 and 2) are proved for general norms. However, his high probability results are proved for Euclidean norms only. In our approach all results hold for general norms.
6. **Simplification.** Our analysis is more compact.

2.3 Assumptions and the Algorithm

Block structure. We model the block structure of the problem by decomposing the space \mathbf{R}^N into n subspaces as follows. Let $U \in \mathbf{R}^{N \times N}$ be a column permutation of the $N \times N$ identity matrix and further let $U = [U_1, U_2, \dots, U_n]$ be a decomposition of U into n submatrices, with U_i being of size $N \times N_i$, where $\sum_i N_i = N$. Clearly, any vector $x \in \mathbf{R}^N$ can be written uniquely as $x = \sum_i U_i x^{(i)}$, where $x^{(i)} = U_i^T x \in \mathbf{R}_i \equiv \mathbf{R}^{N_i}$. Also note that

$$U_i^T U_j = \begin{cases} N_i \times N_i & \text{identity matrix,} & \text{if } i = j, \\ N_i \times N_j & \text{zero matrix,} & \text{otherwise.} \end{cases} \quad (2.2)$$

For simplicity we will write $x = (x^{(1)}, \dots, x^{(n)})^T$. We equip \mathbf{R}_i with a pair of conjugate Euclidean norms:

$$\|t\|_{(i)} = \langle B_i t, t \rangle^{1/2}, \quad \|t\|_{(i)}^* = \langle B_i^{-1} t, t \rangle^{1/2}, \quad t \in \mathbf{R}_i, \quad (2.3)$$

where $B_i \in \mathbf{R}^{N_i \times N_i}$ is a positive definite matrix and $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product.

Example 1. Let $n = N$, $N_i = 1$ for all i and $U = [e_1, e_2, \dots, e_n]$ be the $n \times n$ identity matrix. Then $U_i = e_i$ is the i -th unit vector and $x^{(i)} = e_i^T x \in \mathbf{R}_i = \mathbf{R}$ is the i -th coordinate of x . Also, $x = \sum_i e_i x^{(i)}$. If we let $B_i = 1$ for all i , then $\|t\|_{(i)} = \|t\|_{(i)}^* = |t|$ for all $t \in \mathbf{R}$.

Smoothness of f . We assume throughout the paper that the gradient of f is block coordinate-wise Lipschitz, uniformly in x , with positive constants L_1, \dots, L_n , i.e., that for all $x \in \mathbf{R}^N$, $t \in \mathbf{R}_i$ and i we have

$$\|\nabla_i f(x + U_i t) - \nabla_i f(x)\|_{(i)}^* \leq L_i \|t\|_{(i)}, \quad (2.4)$$

where

$$\nabla_i f(x) \stackrel{\text{def}}{=} (\nabla f(x))^{(i)} = U_i^T \nabla f(x) \in \mathbf{R}_i. \quad (2.5)$$

An important consequence of (2.4) is the following standard inequality [11]:

$$f(x + U_i t) \leq f(x) + \langle \nabla_i f(x), t \rangle + \frac{L_i}{2} \|t\|_{(i)}^2. \quad (2.6)$$

Separability of Ψ . We assume that Ψ is block separable, i.e., that it can be decomposed as follows:

$$\Psi(x) = \sum_{i=1}^n \Psi_i(x^{(i)}), \quad (2.7)$$

where the functions $\Psi_i : \mathbf{R}_i \rightarrow \mathbf{R}$ are convex and closed.

The algorithm. Notice that an upper bound on $F(x+U_it)$, viewed as a function of $t \in \mathbf{R}_i$, is readily available:

$$F(x + U_it) \stackrel{(2.1)}{=} f(x + U_it) + \Psi(x + U_it) \stackrel{(2.6)}{\leq} f(x) + V_i(x, t) + C_i(x), \quad (2.8)$$

where

$$V_i(x, t) \stackrel{\text{def}}{=} \langle \nabla_i f(x), t \rangle + \frac{L_i}{2} \|t\|_{(i)}^2 + \Psi_i(x^{(i)} + t) \quad (2.9)$$

and

$$C_i(x) \stackrel{\text{def}}{=} \sum_{j \neq i} \Psi_j(x^{(j)}). \quad (2.10)$$

We are now ready to describe the generic method. Given iterate x_k , Algorithm 1 picks block $i_k = i \in \{1, 2, \dots, n\}$ with probability $p_i > 0$ and then updates the i -th block of x_k so as to minimize (exactly) in t the upper bound (2.8) on $F(x_k + U_it)$. Note that in certain cases it is possible to minimize $F(x_k + U_it)$ directly; perhaps in a closed form. This is the case, for example, when f is a convex quadratic.

Algorithm 1 RCDC(p, x_0) (**R**andomized **C**oordinate **D**escent for **C**omposite Functions)

for $k = 0, 1, 2, \dots$ **do**

 Choose $i_k = i \in \{1, 2, \dots, n\}$ with probability p_i

$T^{(i)}(x_k) \stackrel{\text{def}}{=} \arg \min \{V_i(x_k, t) : t \in \mathbf{R}_i\}$

$x_{k+1} = x_k + U_i T^{(i)}(x_k)$

end for

The iterates $\{x_k\}$ are random vectors and the values $\{F(x_k)\}$ are random variables. Clearly, x_{k+1} depends only on x_k . As our analysis will be based on the (expected) per-iteration decrease of the objective function, the results will hold even if we replace $V_i(x_k, t)$ by $F(x_k + U_it)$ in Algorithm 1.

Global structure. For fixed positive scalars w_1, \dots, w_n let $W = \text{Diag}(w_1, \dots, w_n)$ and define a pair of conjugate norms in \mathbf{R}^N by

$$\|x\|_W = \left[\sum_{i=1}^n w_i \|x^{(i)}\|_{(i)}^2 \right]^{1/2}, \quad (2.11)$$

$$\|y\|_W^* = \max_{\|x\|_W \leq 1} \langle y, x \rangle = \left[\sum_{i=1}^n w_i^{-1} (\|y^{(i)}\|_{(i)}^*)^2 \right]^{1/2}. \quad (2.12)$$

In the subsequent analysis we will use $W = L$ (Section 2.4), where $L = \text{Diag}(L_1, \dots, L_n)$ and $P = \text{Diag}(p_1, \dots, p_n)$.

The set of optimal solutions of (2.1) is denoted by X^* and x^* is any element of that set. Define

$$\mathcal{R}_W(x) = \max_y \max_{x^* \in X^*} \{\|y - x^*\|_W : F(y) \leq F(x)\},$$

which is a measure of the size of the level set of F given by x . In most of the results in this paper we will need to assume that $\mathcal{R}_W(x_0)$ is finite for the initial iterate x_0 and $W = L$ or $W = LP^{-1}$.

A technical result. The next simple result is the main technical tool enabling us to simplify and improve the corresponding analysis in [13]. It will be used with $\xi_k = F(x_k) - F^*$.

Theorem 1. *Let $\xi_0 > 0$ be a constant, $0 < \epsilon < \xi_0$, and consider a nonnegative nonincreasing sequence of (discrete) random variables $\{\xi_k\}_{k \geq 0}$ with one of the following properties:*

(i) $\mathbf{E}[\xi_{k+1} \mid \xi_k] \leq \xi_k - \frac{\xi_k^2}{c}$, for all k , where $c > 0$ is a constant,

(ii) $\mathbf{E}[\xi_{k+1} \mid \xi_k] \leq (1 - \frac{1}{c})\xi_k$, for all k such that $\xi_k \geq \epsilon$, where $c > 1$ is a constant.

Choose confidence level $\rho \in (0, 1)$. If property (i) holds and we choose $\epsilon < c$ and

$$K \geq \frac{c}{\epsilon}(1 + \log \frac{1}{\rho}) + 2 - \frac{c}{\xi_0}, \quad (2.13)$$

or if property (ii) holds, and we choose

$$K \geq c \log \frac{\xi_0}{\epsilon \rho}, \quad (2.14)$$

then

$$\mathbf{Prob}(\xi_K \leq \epsilon) \geq 1 - \rho. \quad (2.15)$$

Proof. Notice that the sequence $\{\xi_k^\epsilon\}_{k \geq 0}$ defined by

$$\xi_k^\epsilon = \begin{cases} \xi_k & \text{if } \xi_k \geq \epsilon, \\ 0 & \text{otherwise,} \end{cases}$$

satisfies

$$\xi_k^\epsilon \leq \epsilon \iff \xi_k \leq \epsilon, \quad k \geq 0. \quad (2.16)$$

Therefore, by Markov inequality,

$$\mathbf{Prob}(\xi_k > \epsilon) = \mathbf{Prob}(\xi_k^\epsilon > \epsilon) \leq \frac{\mathbf{E}[\xi_k^\epsilon]}{\epsilon},$$

and hence it suffices to show that

$$\theta_K \leq \epsilon \rho, \quad (2.17)$$

where $\theta_k \stackrel{\text{def}}{=} \mathbf{E}[\xi_k^\epsilon]$. If property (i) holds, then

$$\mathbf{E}[\xi_{k+1}^\epsilon \mid \xi_k^\epsilon] \leq \xi_k^\epsilon - \frac{(\xi_k^\epsilon)^2}{c}, \quad \mathbf{E}[\xi_{k+1}^\epsilon \mid \xi_k^\epsilon] \leq (1 - \frac{\epsilon}{c})\xi_k^\epsilon, \quad k \geq 0, \quad (2.18)$$

and by taking expectations (using convexity of $t \mapsto t^2$ in the first case) we obtain

$$\theta_{k+1} \leq \theta_k - \frac{\theta_k^2}{c}, \quad k \geq 0, \quad (2.19)$$

$$\theta_{k+1} \leq (1 - \frac{\epsilon}{c})\theta_k, \quad k \geq 0. \quad (2.20)$$

Notice that (2.19) is better than (2.20) precisely when $\theta_k > \epsilon$. Since

$$\frac{1}{\theta_{k+1}} - \frac{1}{\theta_k} = \frac{\theta_k - \theta_{k+1}}{\theta_{k+1}\theta_k} \geq \frac{\theta_k - \theta_{k+1}}{\theta_k^2} \stackrel{(2.19)}{\geq} \frac{1}{c},$$

we have $\frac{1}{\theta_k} \geq \frac{1}{\theta_0} + \frac{k}{c} = \frac{1}{\xi_0} + \frac{k}{c}$. Therefore, if we let $k_1 \geq \frac{c}{\epsilon} - \frac{c}{\xi_0}$, we obtain $\theta_{k_1} \leq \epsilon$. Finally, letting $k_2 \geq \frac{c}{\epsilon} \log \frac{1}{\rho}$, we have

$$\theta_K \stackrel{(2.13)}{\leq} \theta_{k_1+k_2} \stackrel{(2.20)}{\leq} (1 - \frac{\epsilon}{c})^{k_2} \theta_{k_1} \leq ((1 - \frac{\epsilon}{c})^{\frac{1}{\epsilon}})^{c \log \frac{1}{\rho}} \epsilon \leq (e^{-\frac{1}{c}})^{c \log \frac{1}{\rho}} \epsilon = \epsilon \rho,$$

establishing (2.17). If property (ii) holds, then $\mathbf{E}[\xi_{k+1}^\epsilon \mid \xi_k^\epsilon] \leq (1 - \frac{1}{c})\xi_k^\epsilon$ for all k , and hence

$$\theta_K \leq (1 - \frac{1}{c})^K \theta_0 = (1 - \frac{1}{c})^K \xi_0 \stackrel{(2.14)}{\leq} ((1 - \frac{1}{c})^c)^{\log \frac{\xi_0}{\epsilon \rho}} \xi_0 \leq (e^{-1})^{\log \frac{\xi_0}{\epsilon \rho}} \xi_0 = \epsilon \rho,$$

again establishing (2.17). \square

Tightness. In appendix is shown on simple example that the bounds in the above result are *tight*.

2.4 Coordinate Descent for Composite Functions

In this section we study the performance of Algorithm 1 in the special case when all probabilities are chosen to be the same, i.e., $p_i = \frac{1}{n}$ for all i . For easier future reference we set this method apart and give it a name (Algorithm 2).

Algorithm 2 UCDC(x_0) (Uniform Coordinate Descent for Composite Functions)

for $k = 0, 1, 2, \dots$ **do**

Choose $i_k = i \in \{1, 2, \dots, n\}$ with probability $\frac{1}{n}$

$T^{(i)}(x_k) = \arg \min\{V_i(x_k, t) : t \in \mathbf{R}_i\}$

$x_{k+1} = x_k + U_i T^{(i)}(x_k)$

end for

The following function plays a central role in our analysis:

$$H(x, T) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), T \rangle + \frac{1}{2} \|T\|_L^2 + \Psi(x + T). \quad (2.21)$$

Comparing (2.21) with (2.9) using (2.2), (2.5), (2.7) and (2.11) we get

$$H(x, T) = f(x) + \sum_{i=1}^n V_i(x, T^{(i)}). \quad (2.22)$$

Therefore, the vector $T(x) = (T^{(1)}(x), \dots, T^{(n)}(x))$, with the components $T^{(i)}(x)$ defined in Algorithm 1, is the minimizer of $H(x, \cdot)$:

$$T(x) = \arg \min_{T \in \mathbf{R}^N} H(x, T). \quad (2.23)$$

Let us start by establishing an auxiliary result which will be used repeatedly.

Lemma 1. *Let $\{x_k\}$, $k \geq 0$, be the random iterates generated by $UCDC(x_0)$. Then*

$$\mathbf{E}[F(x_{k+1}) - F^* \mid x_k] \leq \frac{1}{n} (H(x_k, T(x_k)) - F^*) + \frac{n-1}{n} (F(x_k) - F^*). \quad (2.24)$$

Proof.

$$\begin{aligned} \mathbf{E}[F(x_{k+1}) \mid x_k] &= \sum_{i=1}^n \frac{1}{n} F(x_k + U_i T^{(i)}(x_k)) \\ &\stackrel{(2.8)}{\leq} \frac{1}{n} \sum_{i=1}^n [f(x_k) + V_i(x_k, T^{(i)}(x_k)) + C_i(x_k)] \\ &\stackrel{(2.22)}{=} \frac{1}{n} H(x_k, T(x_k)) + \frac{n-1}{n} f(x_k) + \frac{1}{n} \sum_{i=1}^n C_i(x_k) \\ &\stackrel{(2.10)}{=} \frac{1}{n} H(x_k, T(x_k)) + \frac{n-1}{n} f(x_k) + \frac{1}{n} \sum_{i=1}^n \sum_{j \neq i} \Psi_j(x_k^{(j)}) \\ &= \frac{1}{n} H(x_k, T(x_k)) + \frac{n-1}{n} F(x_k). \end{aligned}$$

□

2.4.1 Convex Objective

In order for Lemma 1 to be useful, we need to estimate $H(x_k, T(x_k)) - F^*$ from above in terms of $F(x_k) - F^*$.

Lemma 2. *Fix $x^* \in X^*$, $x \in \text{dom } \Psi$ and let $R = \|x - x^*\|_L$. Then*

$$H(x, T(x)) - F^* \leq \begin{cases} \left(1 - \frac{F(x) - F^*}{2R^2}\right) (F(x) - F^*), & \text{if } F(x) - F^* \leq R^2, \\ \frac{1}{2}R^2 < \frac{1}{2}(F(x) - F^*), & \text{otherwise.} \end{cases} \quad (2.25)$$

Proof.

$$\begin{aligned}
H(x, T(x)) &\stackrel{(2.23)}{=} \min_{T \in \mathbf{R}^N} H(x, T) \\
&= \min_{y \in \mathbf{R}^N} H(x, y - x) \\
&\stackrel{(2.21)}{\leq} \min_{y \in \mathbf{R}^N} f(x) + \langle \nabla f(x), y - x \rangle + \Psi(y) + \frac{1}{2} \|y - x\|_L^2 \\
&\leq \min_{y \in \mathbf{R}^N} F(y) + \frac{1}{2} \|y - x\|_L^2 \\
&\leq \min_{\alpha \in [0, 1]} F(\alpha x^* + (1 - \alpha)x) + \frac{\alpha^2}{2} \|x - x^*\|_L^2 \\
&\leq \min_{\alpha \in [0, 1]} F(x) - \alpha(F(x) - F^*) + \frac{\alpha^2}{2} R^2. \tag{2.26}
\end{aligned}$$

Minimizing (2.26) in α gives $\alpha^* = \min\{1, (F(x) - F^*)/R^2\}$; the result follows. \square

We are now ready to estimate the number of iterations needed to push the objective value within ϵ of the optimal value with high probability. Note that since ρ appears under the logarithm and hence it is easy to attain high confidence.

Theorem 2. *Choose initial point x_0 and target confidence $0 < \rho < 1$. Further, let the target accuracy $\epsilon > 0$ and iteration counter k be chosen in any of the following two ways:*

(i) $\epsilon < F(x_0) - F^*$ and

$$k \geq \frac{2n \max\{\mathcal{R}_L^2(x_0), F(x_0) - F^*\}}{\epsilon} \left(1 + \log \frac{1}{\rho}\right) + 2 - \frac{2n \max\{\mathcal{R}_L^2(x_0), F(x_0) - F^*\}}{F(x_0) - F^*}, \tag{2.27}$$

(ii) $\epsilon < \min\{\mathcal{R}_L^2(x_0), F(x_0) - F^*\}$ and

$$k \geq \frac{2n \mathcal{R}_L^2(x_0)}{\epsilon} \log \frac{F(x_0) - F^*}{\epsilon \rho}. \tag{2.28}$$

If x_k is the random point generated by UCDC(x_0) as applied to the convex function F , then

$$\mathbf{Prob}(F(x_k) - F^* \leq \epsilon) \geq 1 - \rho.$$

Proof. Since $F(x_k) \leq F(x_0)$ for all k , we have $\|x_k - x^*\|_L \leq \mathcal{R}_L(x_0)$ for all $x^* \in X^*$. Lemma 1 together with Lemma 2 then imply that the following holds for all k :

$$\begin{aligned}
\mathbf{E}[F(x_{k+1}) - F^* \mid x_k] &\leq \frac{1}{n} \max \left\{ 1 - \frac{F(x_k) - F^*}{2\|x_k - x^*\|_L^2}, \frac{1}{2} \right\} (F(x_k) - F^*) + \frac{n-1}{n} (F(x_k) - F^*) \\
&= \max \left\{ 1 - \frac{F(x_k) - F^*}{2n\|x_k - x^*\|_L^2}, 1 - \frac{1}{2n} \right\} (F(x_k) - F^*) \\
&\leq \max \left\{ 1 - \frac{F(x_k) - F^*}{2n\mathcal{R}_L^2(x_0)}, 1 - \frac{1}{2n} \right\} (F(x_k) - F^*). \tag{2.29}
\end{aligned}$$

Let $\xi_k = F(x_k) - F^*$ and consider case (i). If we let $c = 2n \max\{\mathcal{R}_L^2(x_0), F(x_0) - F^*\}$, then from (2.29) we obtain

$$\mathbf{E}[\xi_{k+1} \mid \xi_k] \leq \left(1 - \frac{\xi_k}{c}\right)\xi_k = \xi_k - \frac{\xi_k^2}{c}, \quad k \geq 0.$$

Moreover, $\epsilon < \xi_0 < c$. The result then follows by applying Theorem 1. Consider now case (ii). Letting $c = \frac{2n\mathcal{R}_L^2(x_0)}{\epsilon} > 1$, notice that if $\xi_k \geq \epsilon$, inequality (2.29) implies that

$$\mathbf{E}[\xi_{k+1} \mid \xi_k] \leq \max\left\{1 - \frac{\epsilon}{2n\mathcal{R}_L^2(x_0)}, 1 - \frac{1}{2n}\right\}\xi_k = \left(1 - \frac{1}{c}\right)\xi_k.$$

Again, the result follows from Theorem 1. \square

2.4.2 Strongly Convex Objective

Assume that F is strongly convex with respect to some norm $\|\cdot\|$ with convexity parameter $\mu > 0$; that is,

$$F(x) \geq F(y) + \langle F'(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2, \quad x, y \in \text{dom } F, \quad (2.30)$$

where $F'(y)$ is any subgradient of F at y . Note that from the first order optimality conditions for (2.1) we obtain $\langle F'(x^*), x - x^* \rangle \geq 0$ for all $x \in \text{dom } F$ which, combining with (2.30) used with $y = x^*$, yields the standard inequality

$$F(x) - F^* \geq \frac{\mu}{2}\|x - x^*\|^2, \quad x \in \text{dom } F. \quad (2.31)$$

The next lemma will be useful in proving linear convergence of the expected value of the objective function to the minimum.

Lemma 3. *If F is strongly convex with respect to $\|\cdot\|_L$ with convexity parameter $\mu > 0$, then*

$$H(x, T(x)) - F^* \leq \gamma_\mu(F(x) - F^*), \quad x \in \text{dom } F, \quad (2.32)$$

where

$$\gamma_\mu = \begin{cases} 1 - \frac{\mu}{4}, & \text{if } \mu \leq 2, \\ \frac{1}{\mu}, & \text{otherwise.} \end{cases} \quad (2.33)$$

Proof.

$$\begin{aligned}
H(x, T(x)) &\stackrel{(2.23)}{=} \min_{t \in \mathbf{R}^N} H(x, t) \\
&= \min_{y \in \mathbf{R}^N} H(x, y - x) \\
&\leq \min_{y \in \mathbf{R}^N} F(y) + \frac{1}{2} \|y - x\|_L^2 \\
&\leq \min_{\alpha \in [0, 1]} F(\alpha x^* + (1 - \alpha)x) + \frac{\alpha^2}{2} \|x - x^*\|_L^2 \\
&\leq \min_{\alpha \in [0, 1]} F(x) - \alpha(F(x) - F^*) + \frac{\alpha^2}{2} \|x - x^*\|_L^2 \\
&\stackrel{(2.31)}{\leq} \min_{\alpha \in [0, 1]} F(x) + \alpha \left(\frac{\alpha}{\mu} - 1 \right) (F(x) - F^*). \tag{2.34}
\end{aligned}$$

The optimal α in (2.34) is $\alpha^* = \min \{1, \frac{\mu}{2}\}$; the result follows. \square

We now show that the expected value of $F(x_k)$ converges to F^* linearly.

Theorem 3. *Let F be strongly convex with respect to the norm $\|\cdot\|_L$ with convexity parameter $\mu > 0$. If x_k is the random point generated UCDC(x_0), then*

$$\mathbf{E}[F(x_k) - F^*] \leq \left(1 - \frac{1 - \gamma_\mu}{n}\right)^k (F(x_0) - F^*), \tag{2.35}$$

where γ_μ is defined by (2.33).

Proof. Follows from Lemma 1 and Lemma 3. \square

The following is an analogue of Theorem 2 in the case of a strongly convex objective. Note that both the accuracy and confidence parameters appear under the logarithm.

Theorem 4. *Let F be strongly convex with respect to $\|\cdot\|_L$ with convexity parameter $\mu > 0$ and choose accuracy level $\epsilon > 0$, confidence level $0 < \rho < 1$, and*

$$k \geq \frac{n}{1 - \gamma_\mu} \log \left(\frac{F(x_0) - F^*}{\rho \epsilon} \right), \tag{2.36}$$

where γ_μ is given by (2.33). If x_k is the random point generated by UCDC(x_0), then

$$\mathbf{Prob}(F(x_k) - F^* \leq \epsilon) \geq 1 - \rho.$$

Proof. Using Markov inequality and Theorem 3, we obtain

$$\mathbf{Prob}[F(x_k) - F^* \geq \epsilon] \leq \frac{1}{\epsilon} \mathbf{E}[F(x_k) - F^*] \stackrel{(2.35)}{\leq} \frac{1}{\epsilon} \left(1 - \frac{1 - \gamma_\mu}{n}\right)^k (F(x_0) - F^*) \stackrel{(2.36)}{\leq} \rho.$$

\square

Chapter 3

Efficient Serial and Parallel Coordinate Descent Methods for Huge-Scale Truss Topology Design (Abstract)

Abstract In this paper we propose solving *huge-scale* instances of the truss topology design problem with coordinate descent methods. We develop four efficient codes: *serial* and *parallel* implementations of *randomized* and *greedy* rules for the selection of the variable (potential bar) to be updated in the next iteration. Both serial methods enjoy an $O(n/k)$ iteration complexity guarantee, where n is the number of potential bars and k the iteration counter. Our parallel implementations, written in CUDA and running on a graphical processing unit (GPU), are capable of speedups of up to two orders of magnitude when compared to their serial counterparts. Numerical experiments were performed on instances with up to 30 million potential bars.

Bibliography

- [1] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, Sept. 1999.
- [2] A. A. Canutescu and R. L. Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12:963–972, 2003.
- [3] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. Coordinate descent method for large-scale l2-loss linear support vector machines. *Journal of Machine Learning Research*, 9:1369–1398, 2008.
- [4] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. Technical report, 2010.
- [5] J.-B. Hiriart-Urruty and C. Lemaréchal. Fundamentals of convex analysis.
- [6] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear svm. In *In ICML 2008*, pages 408–415, 2008.
- [7] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.
- [8] Y. Li and S. Osher. Coordinate descent optimization for l_1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Probl. Imaging*, 3:487–503, August 2009.
- [9] Z. Q. Luo and P. Tseng. A coordinate gradient descent method for nonsmooth separable minimization. *Journal of optimization theory and applications*, 72(1), January 2002.
- [10] L. Meier, S. V. D. Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society B*, 70:53–71, 2008.
- [11] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Springer Netherlands, 1 edition.
- [12] Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Sept. 2007.

- [13] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. CORE Discussion Paper #2010/2, February 2010.
- [14] Z. T. Qin, K. Scheinberg, and D. Goldfarb. Efficient block-coordinate descent algorithms for the group lasso. Technical report, 2010.
- [15] P. Richtárik and M. Takáč. Efficiency of randomized coordinate descent methods on minimization problems with a composite objective function. Technical report, 2011.
- [16] P. Richtárik and M. Takáč. Efficient serial and parallel coordinate descent method for huge-scale truss topology design. Technical Report ERGO 11-012, 2011.
- [17] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Technical Report ERGO 11-011, 2011.
- [18] A. Saha and A. Tewari. On the finite time convergence of cyclic coordinate descent methods, 2010. preprint arXiv:1005.2146.
- [19] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l1 regularized loss minimization. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.
- [20] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15:262–278, 2009.
- [21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:268–288, 1996.
- [22] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109:475–494, June 2001.
- [23] P. Tseng and S. Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications*, 140:513–535, 2009. 10.1007/s10957-008-9458-3.
- [24] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009. 10.1007/s10107-007-0170-0.
- [25] P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Comput. Optim. Appl.*, 47:179–206, October 2010.
- [26] Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefinite programming. In M. F. Anjos and J. B. Lasserre, editors, *Handbook on Semidefinite, Cone and Polynomial Optimization: Theory, Algorithms, Software and Applications*. Springer, forthcoming.

- [27] S. J. Wright. Accelerated block-coordinate relaxation for regularized optimization. Technical report, 2010.
- [28] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *Trans. Sig. Proc.*, 57:2479–2493, July 2009.
- [29] T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.
- [30] G.-X. Yuan, H. Chia-Hua, and C.-J. Lin. Recent advances of large-scale linear classification. Technical report, 2011.
- [31] G.-X. Yuan and C.-J. Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *Journal of Machine Learning Research*, 11(1):3183–3234, 2010.
- [32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68:49–67, 2006.
- [33] S. Yun and K.-C. Toh. A coordinate gradient descent method for l1-regularized convex minimization. *Computational Optimization and Applications*, 48:273–307, 2011. 10.1007/s10589-009-9251-8.
- [34] H. Zhou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005.