On Optimal Probabilities in Stochastic Coordinate Descent Methods

Peter Richtárik and Martin Takáč
University of Edinburgh, United Kingdom
October 11, 2013

Abstract

We propose and analyze a new parallel coordinate descent method—'NSync—in which at each iteration a random subset of coordinates is updated, in parallel, allowing for the subsets to be chosen *non-uniformly.* We derive convergence rates under a strong convexity assumption, and comment on how to assign probabilities to the sets to optimize the bound. The complexity and practical performance of the method can outperform its uniform variant by an order of magnitude. Surprisingly, the strategy of updating a single randomly selected coordinate per iteration—with optimal probabilities—may require less iterations, both in theory and practice, than the strategy of updating all coordinates at every iteration.

1 Introduction

In this work we consider the optimization problem

$$\min_{x\in\mathbf{R}^n} \phi(x), \tag{1}$$

where ϕ is strongly convex and smooth. We propose a new algorithm, and call it 'NSync (Nonuniform SYNchronous Coordinate descent).

Algorithm 1 ('NSync)

Input: Initial point $x^0 \in \mathbf{R}^n$, subset probabilities $\{p_S\}$ and stepsize parameters $w_1,\dots,w_n > 0$
for $k = 0, 1, 2, \dots$ **do**
  Select a random set of coordinates $\hat{S} \subseteq \{1,\dots,n\}$ such that $\mathbf{Prob}(\hat{S} = S) = p_S$
  Updated selected coordinates: $x^{k+1} = x^k - \sum_{i\in\hat{S}} \frac{1}{w_i}\nabla_i\phi(x^k)e^i$
end for

In 'NSync, we first assign a probability $p_S \geq 0$ to every subset S of $[n] := \{1,\dots,n\}$, with $\sum_S p_S = 1$, and pick stepsize parameters $w_i > 0$, $i = 1, 2, \dots, n$. At every iteration, a random set $\hat{S}$ is generated, independently from previous iterations, following the law $\mathbf{Prob}(\hat{S} = S) = p_S$, and then coordinates $i \in \hat{S}$ are updated in parallel by moving in the direction of the negative partial derivative with stepsize $1/w_i$. The updates are synchronized: no processor/thread is allowed to proceed before all updates are applied, generating the new iterate x^{k+1}. We specifically study samplings $\hat{S}$ which are *non-uniform* in the sense that $p_i := \mathbf{Prob}(i \in \hat{S}) = \sum_{S:i\in S} p_S$ is allowed to vary with i. By $\nabla_i\phi(x)$ we mean $\langle\nabla\phi(x), e^i\rangle$, where $e^i \in \mathbf{R}^n$ is the i-th unit coordinate vector.

Literature. Serial stochastic coordinate descent methods were proposed and analyzed in [6, 13, 15, 18], and more recently in various settings in [12, 7, 8, 9, 21, 19, 24, 3]. Parallel methods were considered in [2, 16, 14], and more recently in [22, 5, 23, 4, 11, 20, 10, 1]. A memory distributed method scaling to big data problems was recently developed in [17]. A nonuniform coordinate

descent method updating a single coordinate at a time was proposed in [15], and one updating two coordinates at a time in [12]. To the best of our knowledge, 'NSync is the first *nonuniform parallel* coordinate descent method.

2 Analysis

Our analysis of 'NSync is based on two assumptions. The first assumption generalizes the ESO concept introduced in [16] and later used in [22, 23, 5, 4, 17] to *nonuniform* samplings. The second assumption requires that ϕ be strongly convex.

Notation: For $x, y, u \in \mathbf{R}^n$ we write $\|x\|_u^2 := \sum_i u_i x_i^2$, $\langle x, y\rangle_u := \sum_{i=1}^n u_i y_i x_i$, $x \bullet y := (x_1 y_1, \ldots, x_n y_n)$ and $u^{-1} := (1/u_1, \ldots, 1/u_n)$. For $S \subseteq [n]$ and $h \in \mathbf{R}^n$, let $h_{[S]} := \sum_{i \in S} h_i e^i$.

Assumption 1 (Nonuniform ESO: Expected Separable Overapproximation)**.** Assume $p = (p_1, \ldots, p_n)^T > 0$ and that for some positive vector $w \in \mathbf{R}^n$ and all $x, h \in \mathbf{R}^n$,

$$\mathbf{E}[\phi(x + h_{[\hat{S}]})] \leq \phi(x) + \langle \nabla \phi(x), h\rangle_p + \tfrac{1}{2}\|h\|_{p \bullet w}^2. \tag{2}$$

Inequalities of type (2), in the *uniform* case ($p_i = p_j$ for all i, j), were studied in [16, 22, 5, 17].

Assumption 2 (Strong convexity)**.** We assume that ϕ is γ-strongly convex with respect to the norm $\|\cdot\|_v$, where $v = (v_1, \ldots, v_n)^T > 0$ and $\gamma > 0$. That is, we require that for all $x, h \in \mathbf{R}^n$,

$$\phi(x + h) \geq \phi(x) + \langle \nabla \phi(x), h\rangle + \tfrac{\gamma}{2}\|h\|_v^2. \tag{3}$$

We can now establish a bound on the number of iterations sufficient for 'NSync to approximately solve (1) with high probability.

Theorem 3. *Let Assumptions 1 and 2 be satisfied. Choose $x^0 \in \mathbf{R}^n$, $0 < \epsilon < \phi(x^0) - \phi^*$ and $0 < \rho < 1$, where $\phi^* := \min_x \phi(x)$. Let*

$$\Lambda := \max_i \tfrac{w_i}{p_i v_i}. \tag{4}$$

If $\{x^k\}$ are the random iterates generated by 'NSync, then

$$K \geq \tfrac{\Lambda}{\gamma} \log\left(\tfrac{\phi(x^0) - \phi^*}{\epsilon\rho}\right) \qquad \Rightarrow \qquad \mathbf{Prob}(\phi(x^K) - \phi^* \leq \epsilon) \geq 1 - \rho. \tag{5}$$

Moreover, we have the lower bound $\Lambda \geq (\sum_i \tfrac{w_i}{v_i})/\mathbf{E}[|\hat{S}|]$.

Proof. We first claim that ϕ is μ-strongly convex with respect to the norm $\|\cdot\|_{w \bullet p^{-1}}$, i.e.,

$$\phi(x + h) \geq \phi(x) + \langle \nabla \phi(x), h\rangle + \tfrac{\mu}{2}\|h\|_{w \bullet p^{-1}}^2, \tag{6}$$

where $\mu := \gamma/\Lambda$. Indeed, this follows by comparing (3) and (6) in the light of (4). Let x^* be such that $\phi(x^*) = \phi^*$. Using (6) with $h = x^* - x$,

$$\phi^* - \phi(x) \overset{(6)}{\geq} \min_{h' \in \mathbf{R}^n} \langle \nabla \phi(x), h'\rangle + \tfrac{\mu}{2}\|h'\|_{w \bullet p^{-1}}^2 = -\tfrac{1}{2\mu}\|\nabla \phi(x)\|_{p \bullet w^{-1}}^2. \tag{7}$$

Let $h^k := -(\mathrm{Diag}(w))^{-1} \nabla \phi(x^k)$. Then $x^{k+1} = x^k + (h^k)_{[\hat{S}]}$, and utilizing Assumption 1, we get

$$\mathbf{E}[\phi(x^{k+1}) \mid x^k] = \mathbf{E}[\phi(x^k + (h^k)_{[\hat{S}]})] \overset{(2)}{\leq} \phi(x^k) + \langle \nabla \phi(x^k), h^k\rangle_p + \tfrac{1}{2}\|h^k\|_{p \bullet w}^2 \tag{8}$$

$$= \phi(x^k) - \tfrac{1}{2}\|\nabla \phi(x^k)\|_{p \bullet w^{-1}}^2 \overset{(7)}{\leq} \phi(x^k) - \mu(\phi(x^k) - \phi^*). \tag{9}$$

Taking expectations in the last inequality and rearranging the terms, we obtain $\mathbf{E}[\phi(x^{k+1}) - \phi^*] \leq (1 - \mu)\mathbf{E}[\phi(x^k) - \phi^*] \leq (1 - \mu)^{k+1}(\phi(x^0) - \phi^*)$. Using this, Markov inequality, and the definition of K, we finally get $\mathbf{Prob}(\phi(x^K) - \phi^* \geq \epsilon) \leq \mathbf{E}[\phi(x^K) - \phi^*]/\epsilon \leq (1 - \mu)^K(\phi(x^0) - \phi^*)/\epsilon \leq \rho$. Let us now establish the last claim. First, note that (see [16, Sec 3.2] for more results of this type),

$$\textstyle\sum_i p_i = \sum_i \sum_{S: i \in S} p_S = \sum_S \sum_{i: i \in S} p_S = \sum_S p_S |S| = \mathbf{E}[|\hat{S}|]. \tag{10}$$

Letting $\Delta := \{p' \in \mathbf{R}^n : p' \geq 0, \sum_i p'_i = \mathbf{E}[|\hat{S}|]\}$, we have

$$\Lambda \overset{(4)+(10)}{\geq} \min_{p' \in \Delta} \max_i \tfrac{w_i}{p'_i v_i} = \tfrac{1}{\mathbf{E}[|\hat{S}|]} \sum_i \tfrac{w_i}{v_i},$$

where the last equality follows since optimal p'_i is proportional to w_i/v_i. □

Theorem 3 is generic in the sense that we do not say when Assumptions 1 and 2 are satisfied, how should one go about to choose the stepsizes w and probabilities $\{p_S\}$. In the next section we address these issues. On the other hand, this abstract setting allowed us to write a brief complexity proof.

Change of variables. Consider the change of variables $y = \mathrm{Diag}(d)x$, where $d > 0$. Defining $\phi^d(y) := \phi(x)$, we get $\nabla\phi^d(y) = (\mathrm{Diag}(d))^{-1}\nabla\phi(x)$. It can be seen that (2), (3) can equivalently be written in terms of ϕ^d, with w replaced by $w^d := w \bullet d^{-2}$ and v replaced by $v^d := v \bullet d^{-2}$. By choosing $d_i = \sqrt{v_i}$, we obtain $v_i^d = 1$ for all i, recovering standard strong convexity.

3 Nonuniform samplings and ESO

Consider now problem (1) with ϕ of the form

$$\phi(x) := f(x) + \tfrac{\gamma}{2}\|x\|_v^2, \tag{11}$$

where $v > 0$. Note that Assumption 2 is satisfied. We further make the following two assumptions.

Assumption 4 (Smoothness)**.** f has Lipschitz gradient with respect to the coordinates, with positive constants $L_1, \dots, L_n$. That is, $|\nabla_i f(x) - \nabla_i f(x + te_i)| \le L_i|t|$ for all $x \in \mathbf{R}^n$ and $t \in \mathbf{R}$.

Assumption 5 (Partial separability)**.** $f(x) = \sum_{J\in\mathcal{J}} f_J(x)$, where $\mathcal{J}$ is a finite collection of nonempty subsets of $[n]$ and f_J are differentiable convex functions such that f_J depends on coordinates $i \in J$ only. Let $\omega := \max_J |J|$. We say that f is *separable of degree* ω.

Uniform parallel coordinate descent methods for regularized problems with f of the above structure were analyzed in [16].

Example 6. Let $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, where $A \in \mathbf{R}^{m\times n}$. Then $L_i = \|A_{:i}\|_2^2$ and $f(x) = \frac{1}{2}\sum_{j=1}^m (A_{j:}x - b_j)^2$, whence ω is the maximum # of nonzeros in a row of A.

Nonuniform sampling. Instead of considering the general case of arbitrary p_S assigned to all subsets of $[n]$, here we consider a special kind of sampling having two advantages: i) sets can be generated easily, ii) it leads to larger stepsizes $1/w_i$ and hence improved convergence rate. Fix $\tau \in [n]$ and $c \ge 1$ and let $S_1, \dots, S_c$ be a collection of (possibly overlapping) subsets of $[n]$ such that $|S_j| \ge \tau$ for all i and $\cup_{j=1}^c S_j = [n]$. Moreover, let $q = (q_1, \dots, q_c) > 0$ be a probability vector. Let $\hat{S}_j$ be τ-nice sampling from S_j; that is, $\hat{S}_j$ picks subsets of S_j having cardinality τ, uniformly at random. We assume these samplings are independent. Now, $\hat{S}$ is generated as follows. We first pick $j \in \{1, \dots, c\}$ with probability q_j, and then draw $\hat{S}_j$. Note that we do not need to compute the quantities p_S, $S \subseteq [n]$, to execute 'NSync. In fact, it is much easier to implement the sampling via the two-tier procedure explained above. Sampling $\hat{S}$ is a nonuniform variant of the τ-nice sampling studied in [16], which here arises as a special case for $c = 1$. Note that

$$p_i = \textstyle\sum_{j=1}^c q_j \frac{\tau}{|S_j|}\delta_{ij} > 0, \quad i \in [n], \tag{12}$$

where $\delta_{ij} = 1$ if $i \in S_j$, and 0 otherwise.

Theorem 7. *Let Assumptions 4 and 5 be satisfied, and let $\hat{S}$ be the sampling described above. Then Assumption 1 is satisfied with p given by* (12) *and any $w = (w_1, \dots, w_n)^T$ for which*

$$w_i \ge w_i^* := \tfrac{L_i + v_i}{p_i} \textstyle\sum_{j=1}^c q_j \frac{\tau}{|S_j|}\delta_{ij}\left(1 + \frac{(\tau-1)(\omega_j - 1)}{\max\{1, |S_j| - 1\}}\right), \qquad i \in [n], \tag{13}$$

where $\omega_j := \max_{J\in\mathcal{J}} |J \cap S_j| \le \omega$.

Proof. Since f is separable of degree ω, so is ϕ (because $\frac{1}{2}\|x\|_v^2$ is separable). Now,

$$\mathbf{E}[\phi(x + h_{[\hat{S}]})] = \mathbf{E}[\mathbf{E}[\phi(x + h_{[\hat{S}_j]}) \mid j]] = \textstyle\sum_{j=1}^c q_j \mathbf{E}[\phi(x + h_{[\hat{S}_j]})] \tag{14}$$

$$\le \textstyle\sum_{j=1}^c q_j \left\{ f(x) + \frac{\tau}{|S_j|}\left(\langle \nabla f(x), h_{[S_j]}\rangle + \frac{1}{2}\left(1 + \frac{(\tau-1)(\omega_j-1)}{\max\{1,|S_j|-1\}}\right)\|h_{[S_j]}\|_{L+v}^2\right)\right\}, \tag{15}$$

where the last inequality follows from the ESO for τ-nice samplings established in [16, Theorem 15]. The claim now follows by comparing the above expression and (2). □

4 Optimal probabilities

Observe that formula (13) can be used to *design* a sampling (characterized by the sets S_j and probabilities q_j) that *minimizes* Λ, which in view of Theorem 3 *optimizes the convergence rate* of the method.

Serial setting. Consider the serial version of 'NSync ($\mathbf{Prob}(|\hat{S}| = 1) = 1$). We can model this via $c = n$, with $S_i = \{i\}$ and $p_i = q_i$ for all $i \in [n]$. In this case, using (12) and (13), we get $w_i = w_i^* = L_i + v_i$. Minimizing Λ in (4) over the probability vector p gives the *optimal probabilities* (we refer to this as the *optimal serial* method) and *optimal complexity*

$$p_i^* = \frac{(L_i+v_i)/v_i}{\sum_j (L_j+v_j)/v_j}, \quad i \in [n], \qquad \Lambda_{OS} = \sum_i \frac{L_i+v_i}{v_i} = n + \sum_i \frac{L_i}{v_i}, \tag{16}$$

respectively. Note that the *uniform sampling*, $p_i = 1/n$ for all i, leads to $\Lambda_{US} := n + n\max_j L_j/v_j$ (we call this the *uniform serial* method), which can be much larger than Λ_{OS}. Moreover, under the change of variables $y = \mathrm{Diag}(d)x$, the gradient of $f^d(y) := f(\mathrm{Diag}(d^{-1})y)$ has coordinate Lipschitz constants $L_i^d = L_i/d_i^2$, while the weights in (11) change to $v_i^d = v_i/d_i^2$. Hence, the condition numbers L_i/v_i can not be improved via such a change of variables.

Optimal serial method can be faster than the fully parallel method. To model the fully parallel setting (i.e., the variant of 'NSync updating *all* coordinates at every iteration), we can set $c = 1$ and $\tau = n$, which yields $\Lambda_{FP} = \omega + \omega\max_j L_j/v_j$. Since $\omega \le n$, it is clear that $\Lambda_{US} \ge \Lambda_{FP}$. However, for large enough ω it will be the case that $\Lambda_{FP} \ge \Lambda_{OS}$, implying, surprisingly, that the optimal serial method can be faster than the fully parallel method.

Parallel setting. Fix τ and sets S_j, $j = 1, 2, \ldots, c$, and define $\theta := \max_j \left(1 + \frac{(\tau-1)(\omega_j-1)}{\max\{1,|S_j|-1\}}\right)$. Consider running 'NSync with stepsizes $w_i = \theta(L_i + v_i)$ (note that $w_i \ge w_i^*$, so we are fine). From (4), (12) and (13) we see that the complexity of 'NSync is determined by

$$\Lambda = \max_i \frac{w_i}{p_i v_i} = \frac{\theta}{\tau}\max_i \left(1 + \frac{L_i}{v_i}\right)\left(\sum_{j=1}^c q_j \frac{\delta_{ij}}{|S_j|}\right)^{-1}.$$

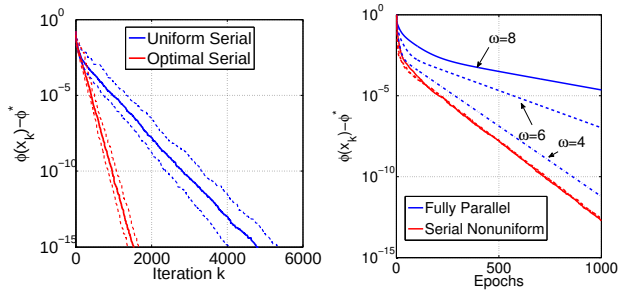
The probability vector q minimizing this quantity can be computed by solving a linear program with $c+1$ variables $(q_1, \ldots, q_c, \alpha)$, $2n$ linear inequality constraints and a single linear equality constraint:

$$\max_{\alpha,q}\left\{\alpha \;\text{ subject to }\; \alpha \le (b^i)^T q \text{ for all } i,\; q \ge 0,\; \sum_j q_j = 1\right\},$$

where $b^i \in \mathbf{R}^c$, $i \in [n]$, are given by $b_j^i = \frac{v_i}{(L_i+v_i)}\frac{\delta_{ij}}{|S_j|}$.

5 Experiments

We now conduct 2 preliminary small scale experiments to illustrate the theory; the results are depicted below. All experiments are with problems of the form (11) with f chosen as in Example 6.



In the **left plot** we chose $A \in \mathbf{R}^{2\times 30}$, $\gamma = 1$, $v_1 = 0.05$, $v_i = 1$ for $i \ne 1$ and $L_i = 1$ for all i. We compare the US method ($p_i = 1/n$, blue) with the OS method (p_i given by (16), red). The dashed lines show 95% confidence intervals (we run the methods 100 times, the line in the middle is the average behavior). While OS can be faster, it is sensitive to over/under-estimation of the constants L_i, v_i. In the **right plot** we show that a nonuniform serial (NS) method can be faster than the fully parallel (FP) variant (we have chosen $m = 8$, $n = 10$ and 3 values of ω). On the horizontal axis we display the number of epochs, where 1 epoch corresponds to updating n coordinates (for FP this is a single iteration, whereas for NS it corresponds to n iterations).

References

[1] Y. Bian, X. Li, and Y. Liu. Parallel coordinate descent Newton for large-scale l1-regularized minimization. *arXiv1306:4080v1*.

[2] J. Bradley, A. Kyrola, D. Bickson, and C. Guestrin. Parallel coordinate descent for l1-regularized loss minimization. In *ICML*, 2011.

[3] C. D. Dang and G. Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. Technical report, Georgia Institute of Technology, 2013.

[4] O. Fercoq. Parallel coordinate descent for the AdaBoost problem. In *ICMLA*, 2013.

[5] O. Fercoq and P. Richtárik. Smooth minimization of nonsmooth functions with parallel coordinate descent methods. *arXiv:1309.5885*, 2013.

[6] C-J. Hsieh, K-W. Chang, C-J. Lin, S.S. Keerthi, , and S. Sundarajan. A dual coordinate descent method for large-scale linear SVM. In *ICML*, 2008.

[7] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletcher. Block-coordinate frank-wolfe optimization for structural svms. In *ICML*, 2013.

[8] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *arXiv:1305.4723*, 2013.

[9] Z. Lu and L. Xiao. Randomized block coordinate non-monotone gradient methods for a class of nonlinear programming. *arXiv:1306.5918*, 2013.

[10] I. Mukherjee, Y. Singer, R. Frongillo, and K. Canini. Parallel boosting with momentum. In *ECML*, 2013.

[11] I. Necoara and D. Clipici. Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: application to distributed mpc. *J. of Process Control*, 23:243–253, 2013.

[12] I. Necoara, Yu. Nesterov, and F. Glineur. Efficiency of randomized coordinate descent methods on optimization problems with linearly coupled constraints. Technical report, 2012.

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

[14] P. Richtárik and M. Takáč. Efficient serial and parallel coordinate descent methods for huge-scale truss topology design. In *Operations Research Proceedings*, pages 27–32. Springer, 2012.

[15] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 2012.

[16] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. *arXiv:1212.0873*, 2012.

[17] P. Richtárik and M. Takáč. Distributed coordinate descent method for learning with big data. *arXiv:1310.2059*, 2013.

[18] S. Shalev-Shwartz and A. Tewari. Stochastic Methods for l1-regularized Loss Minimization. *JMLR*, 12:1865–1892, 2011.

[19] S. Shalev-Shwartz and T. Zhang. Proximal stochastic dual coordinate ascent. *arXiv:1211:2717*, 2012.

[20] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. *arXiv:1305.2581v1*, May 2013.

[21] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *JMLR*, 14:567–599, 2013.

[22] M. Takáč, A. Bijral, P. Richtárik, and N. Srebro. Mini-batch primal and dual methods for SVMs. In *ICML*, 2013.

[23] R. Tappenden, P. Richtárik, and B. Büke. Separable approximations and decomposition methods for the augmented Lagrangian. *arXiv:1308.6774*, 2013.

[24] R. Tappenden, P. Richtárik, and J. Gondzio. Inexact coordinate descent: complexity and preconditioning. *arXiv:1304.5530*, 2013.