On Optimal Probabilities in Stochastic Coordinate Descent Methods

Citation for published version:
Richtárik, P & Taká, M 2016, 'On Optimal Probabilities in Stochastic Coordinate Descent Methods'
Optimization letters, vol 10, no. 6, pp. 1233-1243. DOI: 10.1007/s11590-015-0916-1

Digital Object Identifier (DOI):
10.1007/s11590-015-0916-1

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Optimization letters

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
On optimal probabilities in stochastic coordinate descent methods

Peter Richtárik¹ · Martin Takáč²

Received: 2 January 2015 / Accepted: 16 June 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

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 using an arbitrary probability law. This is the first method of this type. 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.

Keywords Coordinate descent · Arbitrary sampling · First order method · Complexity

1 Introduction

In this work we consider the unconstrained minimization problem

\[
\min_{x \in \mathbb{R}^n} \phi(x),
\]

(1)

¹ University of Edinburgh, Edinburgh, UK
² Lehigh University, Bethlehem, PA, USA
where $\phi$ is strongly convex and differentiable. We propose a new randomized algorithm for solving this problem—NSync (Nonuniform SYNchronous Coordinate descent)—and analyze its iteration complexity. The main novelty of this paper is the algorithm itself. In particular, NSync is the first method which in each iteration updates a random subset of coordinates, allowing for an arbitrary probability law (sampling) to be used for this.

1.1 The algorithm

In NSync (Algorithm 1), we first assign a probability $p_S \geq 0$ to every subset $S$ of the set of coordinates $[n] := \{1, \ldots, n\}$, with

$$\sum_{S \subseteq [n]} p_S = 1,$$

and pick stepsize parameters $w_i > 0$, $i = 1, 2, \ldots, n$, one for each coordinate.

**Algorithm 1** (NSync)

```
Input: Initial point $x^0 \in \mathbb{R}^n$, subset probabilities \{\(p_S\)\} and stepsize parameters \(w_1, \ldots, w_n > 0\) for $k = 0, 1, 2, \ldots$ do
  Select a random set of coordinates $\hat{S} \subseteq \{1, \ldots, n\}$ such that $\text{Prob}(\hat{S} = S) = p_S$
  Update 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
```

At every iteration, a random set $\hat{S}$ is generated, independently from previous iterations, following the law

$$\text{Prob}(\hat{S} = S) = p_S, \quad S \subseteq [n],$$

and then coordinates $i \in \hat{S}$ are updated in parallel by moving in the direction of the negative partial derivative with stepsize $1/w_i$. By $\nabla_i \phi(x)$ we mean $\langle \nabla \phi(x), e^i \rangle$, where $e^i \in \mathbb{R}^n$ is the $i$th unit coordinate vector.

The updates are synchronized: no processor/thread is allowed to proceed before all updates are applied, generating the new iterate $x^{k+1}$. We study the complexity of NSync for arbitrary sampling $\hat{S}$. In particular, $\hat{S}$ can be non-uniform in the sense that the probability that coordinate $i$ is chosen,

$$p_i := \text{Prob}(i \in \hat{S}) = \sum_{S : i \in S} p_S,$$

is allowed to vary with $i$.

1.2 Literature

Serial stochastic coordinate descent methods were proposed and analyzed in [8, 15, 20, 23], and more recently in various settings in [4, 9–11, 14, 24, 26, 29]. Parallel
methods were considered in [2, 19, 21], and more recently in [1, 5, 6, 12, 13, 25, 27, 28]. A memory distributed method scaling to big data problems was recently developed in [22]. A nonuniform coordinate descent method updating a single coordinate at a time was proposed in [20], and one updating two coordinates at a time in [14].

NSync is the first randomized method in the literature which is capable of updating a subset of the coordinates without any restrictions, i.e., according to an arbitrary probability law, except for the necessary requirement that \( p_i > 0 \) for all \( i \). In particular, NSync is the first nonuniform parallel coordinate descent method.

In the time between the first online appearance of this work on arXiv (October 2013; arXiv:1310.3438), and the time this paper went to press, this work led to a number of extensions [3, 7, 16–18]. All of these papers share the defining feature of NSync, namely, its ability to work with an arbitrary probability law defining the selection of the active coordinates in each iteration. These works also utilize the nonuniform ESO assumption introduced here (Assumption 1), as it appears to be key in the study of such methods.

2 Analysis

In this section we provide a complexity analysis of NSync.

2.1 Assumptions

Our analysis of NSync is based on two assumptions. The first assumption generalizes the ESO concept introduced in [21] and later used in [5, 6, 22, 27, 28] to nonuniform samplings. The second assumption requires that \( \phi \) be strongly convex.

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

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

\[
E[\phi(x + h_{[\hat{S}]}))] \leq \phi(x) + (\nabla \phi(x), h)_p + \frac{1}{2} \|h\|_{p \cdot w}^2. \tag{2}
\]

As soon as \( \phi \) has a Lipschitz continuous gradient, then for every random sampling \( \hat{S} \) there exist positive weights \( w_1, \ldots, w_n \) such that Assumption 1 holds. In this sense, the assumption is not restrictive. Inequalities of the type (2), in the uniform case \( (p_i = p_j \text{ for all } i, j) \), were studied in [6, 21, 22, 27]. Motivated by the introduction of the nonuniform ESO assumption in this paper, and the development in Sect. 3 of our
work, an entire paper was recently written, dedicated to the study of nonuniform ESO inequalities \[16\].

We now turn to the second and final assumption.

**Assumption 2** (Strong convexity) We assume that \( \phi \) is \( \gamma \)-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 \( x, h \in \mathbb{R}^n \),

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

**2.2 Complexity**

We can now establish a bound on the number of iterations sufficient for NSync to approximately solve \((1)\) with high probability. We believe it is remarkable that the proof is very concise.

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

\[
\Lambda := \max_i \frac{w_i}{p_i v_i}. \tag{4}
\]

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

\[
K \geq \frac{\Lambda}{\gamma} \log \left( \frac{\phi(x^0) - \phi^*}{\epsilon \rho} \right) \quad \Rightarrow \quad \text{Prob}(\phi(x^K) - \phi^* \leq \epsilon) \geq 1 - \rho. \tag{5}
\]

Moreover, we have the lower bound

\[
\Lambda \geq \left( \sum_{i=1}^{n} \frac{w_i}{v_i} \right) / \mathbb{E}[|\hat{S}|]. \tag{6}
\]

**Proof** We first claim that \( \phi \) is \( \mu \)-strongly convex with respect to the norm \( \| \cdot \|_{w^* p^{-1}} \), i.e.,

\[
\phi(x + h) \geq \phi(x) + \langle \nabla \phi(x), h \rangle + \frac{\mu}{2} \| h \|_{w^* p^{-1}}^2. \tag{7}
\]

\(^1\) A clarifying comment answering a question raised by the reviewer: The authors of \[16\] give explicit formulas for \( w \) for which \((2)\) holds, under an assumption that is slightly weaker than Lipschitz continuity of the gradient of \( \phi \). In particular, they study functions \( \phi \) admitting the global quadratic upper bound

\[
\phi(x + h) \leq \phi(x) + \langle \nabla \phi(x), h \rangle + \frac{1}{2} \| Ah \|_2^2
\]

for all \( x, h \in \mathbb{R}^n \), where \( A \in \mathbb{R}^{m \times n} \). One of the consequence of their work is that the parameters \( w_1, \ldots, w_n \) must necessarily satisfy the inequalities: \( w_i \geq \| A_{i} \|_2^2 \), where \( A_{i} \) is the \( i \)th column of \( A \). Moreover, as long as \( \text{Prob}(|\hat{S}| \leq \tau) = 1 \) for some \( \tau \), then \((2)\) holds for \( w_i = \tau \| A_{i} \|_2^2 \). However, this choice of parameters is rather conservative. The goal of \[16\] is to give explicit and tight formulas for \( w \), where hopefully \( w_i \) will be much smaller than \( \tau \| A_{i} \|_2^2 \), utilizing specific properties of the sampling \( \hat{S} \) and data matrix \( A \).
where $\mu := \gamma / \Lambda$. Indeed, this follows by comparing (3) and (7) in the light of (4). Let $x^*$ be such that $\phi(x^*) = \phi^*$. Using (7) with $h = x^* - x$,

$$
\phi^* - \phi(x) \geq \min_{h' \in \mathbb{R}^n} \langle \nabla \phi(x), h' \rangle + \frac{\mu}{2} \| h' \|_w^2 \geq - \frac{1}{2\mu} \| \nabla \phi(x) \|^2_{p \cdot w^{-1}}.
$$

(8)

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

$$
\mathbb{E}[\phi(x^{k+1}) \mid x^k] = \mathbb{E}[\phi(x^k + (h^k)[\hat{S}])]
\leq \phi(x^k) + \langle \nabla \phi(x^k), h^k \rangle_p + \frac{1}{2} \| h^k \|^2_{p \cdot w}
= \phi(x^k) - \frac{1}{2} \| \nabla \phi(x^k) \|^2_{p \cdot w^{-1}}
\leq \phi(x^k) - \mu (\phi(x^k) - \phi^*) .
$$

Taking expectations in the last inequality and rearranging the terms, we obtain

$$
\mathbb{E}[\phi(x^{k+1}) - \phi^*] \leq (1 - \mu) \mathbb{E}[\phi(x^k) - \phi^*] \leq (1 - \mu)^{k+1} (\phi(x^0) - \phi^*).
$$

Using this, Markov inequality, and the definition of $K$, we finally get

$$
\text{Prob}(\phi(x^K) - \phi^* \geq \epsilon) \leq \frac{\mathbb{E}[\phi(x^K) - \phi^*]}{\epsilon} \leq \frac{(1 - \mu)^K (\phi(x^0) - \phi^*)}{\epsilon} \leq \rho.
$$

Let us now establish the last claim.

First, note that (see [21, Sec 3.2] for more results of this type),

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

(9)

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

$$
\Lambda \geq \min_{p' \in \Delta} \max_i \frac{w_i}{p'_i v_i} = \frac{1}{\mathbb{E}[|\hat{S}|]} \sum_{i=1}^n \frac{v_i}{w_i},
$$

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

Theorem 3 is generic in the sense that we do not say when Assumption 1 is satisfied and how should one go about choosing the stepsizes $\{w_i\}$ 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.

The quantity $\Lambda$, defined in (4), can be interpreted as a condition number associated with the problem and our method. Hence, as we vary the distribution of $\hat{S}$, $\Lambda$ will vary.
It is clear intuitively that $\Lambda$ can be arbitrarily bad. Indeed, by choosing a sampling $\hat{S}$ which “nearly” ignores one or more of the coordinates (by setting $p_i \approx 0$ for some $i$), we should expect the number of iterations to grow as the method will necessarily be very slow in updating these coordinates.

In the light of this, inequality (6) is useful as it gives a useful expression for bounding $\Lambda$ from below.

### 2.3 Change of variables

Consider the change of variables $y = \text{Diag}(d)x$, where $d > 0$. Defining $\phi^d(y) := \phi(x)$, we get $\nabla \phi^d(y) = (\text{Diag}(d))^{-1}\nabla \phi(x)$. It can be seen that (2), (3) can equivalently be written in terms of $\phi^d$, with $w$ replaced by $w^d := w \cdot d^{-2}$ and $v$ replaced by $v^d := v \cdot d^{-2}$. 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

In this section we consider a problem with standard assumptions and show that the (admittedly nonstandard) ESO assumption, Assumption 1, is satisfied.

Consider now problem (1) with $\phi$ of the form

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

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

**Assumption 4** (Smoothness) Function $f$ has Lipschitz gradient with respect to the coordinates, with positive constants $L_1, \ldots, L_n$. That is,

$$
|\nabla_i f(x) - \nabla_i f(x + te_i)| \leq L_i |t|
$$

for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$.

**Assumption 5** (Partial separability) Function $f$ has the form

$$
f(x) = \sum_{J \in \mathcal{J}} f_J(x),
$$

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 $\omega$.

Uniform parallel coordinate descent methods for regularized problems with $f$ of the above structure were analyzed in [21].
Example 1 Let

\[ f(x) = \frac{1}{2} \|Ax - b\|^2, \]

where \( A \in \mathbb{R}^{m \times n} \). Then \( L_i = \|A_i\|^2 \) and

\[ f(x) = \frac{1}{2} \sum_{j=1}^{m} (A_{ij} x - b_j)^2, \]

where \( A_{ij} \) is the \( i \)th column of \( A \), \( A_{j} \) is the \( j \)th row of \( A \) and \( \| \cdot \| \) is the standard L2 norm. Then \( \omega \) 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 \( c \geq 1 \) and let \( S_1, \ldots, S_c \) be a collection of (possibly overlapping) subsets of \([n]\) such that

\[ |S_j| \geq \tau \]

for all \( j = 1, 2, \ldots, c \) and

\[ \bigcup_{j=1}^{c} S_j = [n]. \]

Moreover, let \( q = (q_1, \ldots, q_c) > 0 \) be a probability vector. Let \( \hat{S}_j \) be \( \tau \)-nice sampling from \( S_j \); that is, \( \hat{S}_j \) picks subsets of \( S_j \) having cardinality \( \tau \), uniformly at random. We assume these samplings are independent. Now, \( \hat{S} \) is defined as follows: We first pick \( j \in \{1, \ldots, c\} \) 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 \( \tau \)-nice sampling studied in [21], which here arises as a special case for \( c = 1 \).

Note that

\[ p_i = \sum_{j=1}^{c} q_j \frac{\tau}{|S_j|} \delta_{ij} > 0, \quad i \in [n], \quad (11) \]

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

In our next result we show that Assumption 1 is satisfied for \( f \) and the sampling described above.

Theorem 6 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 (11) and any \( w = (w_1, \ldots, w_n)^T \) for which
\[ w_i \geq w_i^* := \frac{L_i + v_i}{p_i} \sum_{j=1}^{c} q_j \frac{\tau}{|S_j|} \delta_{ij} \left( 1 + \frac{(\tau - 1)(\omega_j - 1)}{\max\{1, |S_j| - 1\}} \right) \]  

(12)

for all \( i \in [n] \), where

\[ \omega_j := \max_{J \in \mathcal{J}} |J \cap S_j| \leq \omega. \]

**Proof** Since \( f \) is separable of degree \( \omega \), so is \( \phi \) (because \( \frac{1}{2} \|x\|_2^2 \) is separable). Now,

\[
\mathbb{E}[\phi(x + h_{\hat{S}_j})] = \mathbb{E}[\mathbb{E}[\phi(x + h_{\hat{S}_j}) | j]] = \sum_{j=1}^{c} q_j \mathbb{E}[\phi(x + h_{\hat{S}_j})] \\
\leq \sum_{j=1}^{c} q_j \left\{ f(x) + \frac{\tau}{|S_j|} \left( \langle \nabla f(x), h_{[S_j]} \rangle + \frac{1}{2} \left( 1 + \frac{(\tau - 1)(\omega_j - 1)}{\max\{1, |S_j| - 1\}} \right) \|h_{[S_j]}\|_{L^2} \right) \right\},
\]

where the last inequality follows from the ESO for \( \tau \)-nice samplings established in [21, Theorem 15]. The claim now follows by comparing the above expression and (2). \( \square \)

### 4 Optimal probabilities

Observe that the formula (12) can be used to design a sampling (characterized by the sets \( S_j \) and probabilities \( q_j \)) that maximizes \( \mu \), which in view of Theorem 3 optimizes the convergence rate of the method.

#### 4.1 Serial setting

Consider the serial version of NSync (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 (11) and (12), we get \( w_i = w_i^* = L_i + v_i \). Minimizing \( \Lambda \) in (4) over the probability vector \( p \) gives the optimal probabilities (we refer to this as the optimal serial method)

\[
p_i^* = \frac{(L_i + v_i)/v_i}{\sum_j (L_j + v_j)/v_j}, \quad i \in [n], \tag{13}
\]

and optimal complexity

\[
\Lambda_{OS} = \sum_{i=1}^{n} \frac{L_i + v_i}{v_i} = n + \sum_{i=1}^{n} \frac{L_i}{v_i}, \tag{14}
\]

Note that the uniform sampling, defined by \( p_i = 1/n \) for all \( i \in [n] \), leads to

\[
\Lambda_{US} := n + n \max_j \frac{L_j}{v_j}.
\]
On optimal probabilities in stochastic...

Note that this can be much larger than $\Lambda_{OS}$. We refer to NSync utilizing this sampling as the uniform serial method.

Moreover, the condition numbers $L_i/v_i$ can not be improved via such a change of variables. Indeed, under the change of variables $y = \text{Diag}(d)x$, the gradient of $f^d(y) := f(\text{Diag}(d^{-1})y)$ has coordinate Lipschitz constants $L_i^d = L_i/d_i^2$, while the weights in (10) change to $v_i^d = v_i/d_i^2$.

4.2 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

$$\Lambda_{FP} = \omega + \omega \max_j \frac{L_j}{v_j}.$$  

Since $\omega \leq n$, it is clear that $\Lambda_{US} \geq \Lambda_{FP}$. However, for large enough $\omega$ it will be the case that $\Lambda_{FP} \geq \Lambda_{OS}$, implying, surprisingly, that the optimal serial method can be faster than the fully parallel method.

4.3 Parallel setting

Fix $\tau$ and set $S_j, j = 1, 2, \ldots, c$, 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 $w_i \geq w_i^*$, so we are fine). From (4), (11) and (12) we see that the complexity of NSync is determined by

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

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:

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

where $b^i \in \mathbb{R}^c, i \in [n]$, are given by

$$b^i_j = \frac{v_i}{(L_i + v_i)} \frac{\delta_{ij}}{|S_j|}.$$
5 Experiments

We now conduct two preliminary small scale experiments to illustrate the theory; the results are depicted in Fig. 1. All experiments are with problems of the form (10) with $f$ chosen as in Example 1.

In the left plot we chose $A \in \mathbb{R}^{2 \times 30}$, $\gamma = 1$, $v_1 = 0.05$, $v_i = 1$ for $i \neq 1$ 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 (13), 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 three values of $\omega$). On the horizontal axis we display the number of epochs, where one epoch corresponds to updating $n$ coordinates (for FP this is a single iteration, whereas for NS it corresponds to $n$ iterations).

Acknowledgments This work appeared on arXiv in October 2013 (arXiv:1310.3438). P. Richtárik and M. Takáč were partially supported by the Centre for Numerical Algorithms and Intelligent Software (funded by EPSRC grant EP/G036136/1 and the Scottish Funding Council). The second author also acknowledge support from the EPSRC Grant EP/K02325X/1, Accelerated Coordinate Descent Methods for Big Data Optimization.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Bian, Y., Li, X., Liu, Y.: Parallel coordinate descent Newton for large-scale l1-regularized minimization. arXiv:1306.4080v1 (2013)
2. Bradley, J., Kyroala, A., Bickson, D., Guestrin, C.: Parallel coordinate descent for L1-regularized loss minimization. In: International Conference on Machine Learning (2011)
3. Csiba, D., Richtárik, P.: Primal method for ERM with flexible mini-batching schemes and non-convex losses. arXiv:1506.02227 (2015)
On optimal probabilities in stochastic...