Parallel (Block) Coordinate Descent Methods for Big Data Optimization
by Peter Richtarik and Martin Takac

Ming Yan

Rice CAAM
Introduction

Parallel Block Coordinate Descent Methods

Block Sampling

Expected Separable Overapproximation (ESO)

Expected Separable Overapproximation (ESO) of Partially Separable Functions

Iteration Complexity
Outline

Introduction

Parallel Block Coordinate Descent Methods

Block Sampling

Expected Separable Overapproximation (ESO)

Expected Separable Overapproximation (ESO) of Partially Separable Functions

Iteration Complexity
Introduction

- **Big data optimization**: The size of problems grows with our capacity to solve them, and is projected to grow dramatically in the next decade. Developing optimization algorithms suitable for the task.

- **Coordinate descent methods**: Coordinate descent methods (CDM) drastically reduces memory requirements as well as the arithmetic complexity of a single iteration, making the methods easily implementable and scalable. As observed by numerous authors, serial CDMs are much more efficient for big data optimization problems than most other competing approaches, such as gradient methods.

- **Parallelization**: For truly huge-scale problems it is absolutely necessary to parallelize.
Problem

Minimizing a partially separable composite objective.

\[
\minimize \; f(x) + \Omega(x) \quad \text{subject to} \quad x \in \mathbb{R}^N,
\]

where \( f \) is a (block) partially separable smooth convex function and \( \Omega \) is a simple (block) separable convex function. Assume that \( \Omega \) is proper and closed, and this problem has a minimum \( (F^* > -\infty) \).
Partial separability

Let $x \in \mathbb{R}^N$ be decomposed into $n$ non-overlapping blocks of variables $x^{(1)}, \ldots, x^{(n)}$. We assume that $f : \mathbb{R}^N \to \mathbb{R}$ is partially separable of degree $\omega$, i.e., that it can be written in the form

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

(1)

where $\mathcal{J}$ is a finite collection of nonempty subsets of $[n] \equiv \{1, 2, \ldots, n\}$ (possibly containing identical sets multiple times), $f_J$ are differentiable convex functions such that $f_J$ depends on blocks $x^{(i)}$ for $i \in J$ only, and

$$|J| \leq \omega \text{ for all } J \in \mathcal{J}.$$

We do not need to know the decomposition, all we need is $\omega$, which is often an easily computable quantity.
Examples of partially separable functions

For simplicity, we assume all blocks are of size 1 (i.e., $N = n$). Let

$$f(x) = \sum_{j=1}^{m} \mathcal{L}(x, A_j, y_j),$$

where $m$ is the number of examples, $x \in \mathbb{R}^n$ is the vector of features, $(A_j, y_j) \in \mathbb{R}^n \times \mathbb{R}$ are labeled examples and $\mathcal{L}$ is one of the three loss functions listed below. Let $A \in \mathbb{R}^{m \times n}$ with row $j$ equal to $A_j^T$. Often, each example depends on a few features only; the maximum over all features is the degree of partial separability $\omega$.

<table>
<thead>
<tr>
<th>Loss Function</th>
<th>Formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square Loss</td>
<td>$\frac{1}{2} (A_j^T x - y_j)^2$</td>
</tr>
<tr>
<td>Logistic Loss</td>
<td>$\log(1 + e^{-y_j A_j^T x})$</td>
</tr>
<tr>
<td>Hinge Square Loss</td>
<td>$\frac{1}{2} \max{0, 1 - y_j A_j^T x}^2$</td>
</tr>
</tbody>
</table>

Table 1: Three examples of loss of functions
Outline

Introduction

Parallel Block Coordinate Descent Methods

Block Sampling

Expected Separable Overapproximation (ESO)

Expected Separable Overapproximation (ESO) of Partially Separable Functions

Iteration Complexity
Let $U \in \mathbb{R}^{N \times N}$ be a column permutation of the $N \times N$ identity matrix and further let $U = [U_1, U_2, \ldots, U_n]$ be a decomposition of $U$ into $n$ submatrices, with $U_i$ being of size $N \times N_i$, where $\sum_i N_i = N$.

- **Block decomposition:** Any vector $x \in \mathbb{R}^N$ can be written uniquely as

$$x = \sum_{i=1}^{n} U_i x^{(i)},$$

where $x^{(i)} \in \mathbb{R}_i \equiv \mathbb{R}^{N_i}$. Moreover $x^{(i)} = U_i^T x$. For now on we write $x^{(i)} \equiv U_i^T x \in \mathbb{R}_i$, and refer to $x^{(i)}$ as the $i$-th block of $x$.

- **Projection onto a set of blocks:** For $S \subset [n]$ and $x \in \mathbb{R}^N$, we write

$$x[S] \equiv \sum_{i \in S} U_i x^{(i)}.$$
Notations

- **Inner products:** Let \( \langle \cdot, \cdot \rangle \) denote the standard Euclidean inner product in space \( \mathbb{R}^N \) and \( \mathbb{R}_i, \ i \in [n] \). \( \langle x, y \rangle = \sum_{i=1}^{n} \langle x^{(i)}, y^{(i)} \rangle \). For any \( w \in \mathbb{R}^n \) and \( x, y \in \mathbb{R}^N \), we define

\[
\langle x, y \rangle \equiv \sum_{i=1}^{n} w_i \langle x^{(i)}, y^{(i)} \rangle.
\]

For vectors \( z = (z_1, \ldots, z_n)^T \) and \( w = (w_1, \ldots, w_n)^T \), we write \( w \odot z \equiv (w_1z_1, \ldots, w_nz_n)^T \).

- **Norms:** Spaces \( \mathbb{R}_i, \ i \in [n] \) are equipped with a pair of conjugate norms: \( \| t \|_{(i)} \) and \( \| t \|_{(i)}^* \equiv \max_{\| s \|_{(i)} \leq 1} \langle s, t \rangle, \ t \in \mathbb{R}_i \). For \( w \in \mathbb{R}_+^n \) define a pair of conjugate norms in \( \mathbb{R}^N \) by

\[
\| x \|_w = \left[ \sum_{i=1}^{n} w_i \| x^{(i)} \|_{(i)}^2 \right]^{1/2}, \quad \| y \|_w^* = \left[ \sum_{i=1}^{n} w_i^{-1} (\| y^{(i)} \|_{(i)}^*)^2 \right]^{1/2}.
\]

We will use \( w = L \equiv (L_1, L_2, \ldots, L_n)^T \), where \( L_i \) are defined below.
Assumptions

- **Smoothness of** $f$: We assume that the gradient of $f$ is block Lipschitz, uniformly in $x$, with positive constants $L_1, \ldots, L_n$, i.e., that for all $x \in \mathbb{R}^N$, $i \in [n]$ and $t \in \mathbb{R}_i$,

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

where $\nabla_i f(x) \equiv (\nabla f(x))^{(i)} = U_i^T \nabla f(x) \in \mathbb{R}_i$. We have

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

- **Separability of** $\Omega$: We assume that $\Omega : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is (block) separable, i.e., that it can be decomposed as follows:

$$
\Omega(x) = \sum_{i=1}^{n} \Omega_i(x^{(i)}),
$$

where the functions $\Omega_i : \mathbb{R}_i \to \mathbb{R} \cup \{+\infty\}$ are convex and closed.
- **Strong convexity:** We will assume that either $f$ or $\Omega$ (or both) is strongly convex. A function $\phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is strongly convex with respect to the norm $\| \cdot \|_w$ with convexity parameter $\mu_{\phi}(w) \geq 0$ if for all $x, y \in \text{dom } \phi$,

$$
\phi(y) \geq \phi(x) + \langle \phi'(x), y - x \rangle + \frac{\mu_{\phi}(w)}{2} \| y - x \|^2_w,
$$

where $\phi'(x)$ is any subgradient of $\phi$ at $x$. The case with $\mu_{\phi}(w) = 0$ reduces to convexity.

$$
\mu_F(w) \geq \mu_f(w) + \mu_{\Omega}(w).
$$

$$
\phi(\lambda x + (1 - \lambda)y) \leq \lambda \phi(x) + (1 - \lambda)\phi(y) - \frac{\mu_{\phi}(w)\lambda(1 - \lambda)}{2} \| x - y \|^2_w.
$$
Algorithms

Algorithm 1 Parallel Coordinate Descent Method 1 (PCDM1)

1: Choose initial point $x_0 \in \mathbb{R}^N$
2: for $k = 0, 1, 2, \ldots$ do
3: Randomly generate a set of blocks $S_k \subset \{1, 2, \ldots, n\}$
4: $x_{k+1} \leftarrow x_k + (h(x_k))[S_k]$
5: end for

Algorithm 2 Parallel Coordinate Descent Method 2 (PCDM2)

1: Choose initial point $x_0 \in \mathbb{R}^N$
2: for $k = 0, 1, 2, \ldots$ do
3: Randomly generate a set of blocks $S_k \subset \{1, 2, \ldots, n\}$
4: $x_{k+1} \leftarrow x_k + (h(x_k))[S_k]$
5: If $F(x_{k+1}) > F(x_k)$, then $x_{k+1} \leftarrow x_k$
6: end for
Step 3: We pick a random set \((S_k)\) of blocks to be updated (in parallel) during that iteration; \(S_k\) is a realization of a random set-valued mapping \(\hat{S}\) with values in \(2^{[n]}\). For brevity, we refer to such a mapping by the name sampling and we limit our attention to uniform samplings, i.e., random sets having the following property: \(P(i \in \hat{S}) = \text{const}\) for all blocks \(i\).

Step 4: For \(x \in \mathbb{R}^N\) we define

\[
h(x) \equiv \arg \min_{h \in \mathbb{R}^N} H_{\beta, w}(x, h),
\]

where

\[
H_{\beta, w}(x, h) \equiv f(x) + \langle \nabla f(x), h \rangle + \frac{\beta}{2} \|h\|_w^2 + \Omega(x + h),
\]

and \(\beta > 0\), and \(w = (w_1, \ldots, w_n)^T \in \mathbb{R}_+^n\) are parameters that will be commented on later. \(H_{\beta, w}(x, \cdot)\) is block separable, and step 4 becomes:

In parallel for \(i \in S_k\) do: \(x_{k+1}^{(i)} \leftarrow x_k^{(i)} + h^{(i)}(x_k)\).
\[
\mathbb{E}[F(x + h_{\hat{S}})] = \mathbb{E}[f(x + h_{\hat{S}}) + \Omega(x + h_{\hat{S}})] \\
\leq f(x) + \frac{\mathbb{E}[|\hat{S}|]}{n} \left( \langle \nabla f(x), h \rangle + \frac{\beta}{2} \| h \|_w^2 \right) \\
+ \left(1 - \frac{\mathbb{E}[|\hat{S}|]}{n}\right) \Omega(x) + \frac{\mathbb{E}[|\hat{S}|]}{n} \Omega(x + h) \\
= \left(1 - \frac{\mathbb{E}[|\hat{S}|]}{n}\right) F(x) + \frac{\mathbb{E}[|\hat{S}|]}{n} H_{\beta,w}(x, h) \\
H_{\beta,w}(x, 0) = F(x)
\]
Step 5: In some situations we are not able to analyze the iteration complexity of PCDM1 (non-strongly-convex $F$ where monotonicity of the method is not guaranteed by other means than by directly enforcing it by inclusion of Step 5).

Let us remark that this issue arises for general $\Omega$ only. It does not exist for $\Omega = 0$, $\Omega(\cdot) = \lambda \| \cdot \|_1$ and for $\Omega$ encoding simple constraints on individual blocks; in these cases one does not need to consider PCDM2.
Contributions

- **Problem generality.** We give the first complexity analysis for a parallel coordinate descent method in its full generality.

- **Partial separability.** We give the first analysis of a coordinate descent type method dealing with a partially separable objective. In order to run the method, all we need to know about $f$ are the Lipschitz constants $L_i$ and the degree of partial separability $\omega$.

- **Algorithm unification.** Depending on the choice of the block structure and the way blocks are selected at every iteration, we give the first analysis of a method which continuously interpolates between a serial coordinate descent method and the gradient method.
- **Expected Separable Overapproximation (ESO).** En route to proving the iteration complexity results for our algorithms, we develop a theory of deterministic and expected separable overapproximation. We believe this is of independent interest; for instance, methods based on ESO can be compared favorably to the Diagonal Quadratic Approximation (DQA) approach used in the decomposition of stochastic optimization programs.

- **Variable number of updates per iteration.** We give the first analysis of a PCDM which allows for a variable number of blocks to be updated throughout the iterations. This may be useful in some settings such as when the problem is being solved in parallel by \( \tau \) unreliable processors each of which computes its update \( h(i)(x_k) \) with probability \( p_b \) and is busy/down with probability \( 1 - p_b \).

- **Parallelization speedup.** We show theoretically and numerically that PCDM accelerates on its serial counterpart for partially separable problems.
Outline

Introduction

Parallel Block Coordinate Descent Methods

Block Sampling

Expected Separable Overapproximation (ESO)

Expected Separable Overapproximation (ESO) of Partially Separable Functions

Iteration Complexity
Block sampling

A sampling $\hat{S}$ is uniquely characterized by the probability density function

$$P(S) \equiv P(\hat{S} = S), \quad S \subset [n].$$

Further we let $p = (p_1, \ldots, p_n)^T$, where

$$p_i \equiv P(i \in \hat{S}).$$

- A sampling is **proper** if $p_i > 0$ for all blocks $i$. PCDM can not converge for a sampling that is not proper.
- A sampling $\hat{S}$ is **uniform** if all blocks get updated with the same probability, i.e., if $p_i = \text{const}$.
- A sampling $\hat{S}$ is **nil** if $P(\emptyset) = 1$. 
Special classes of uniform samplings

- **Doubly uniform (DU) samplings:** A DU sampling is one which generates all sets of equal cardinality with equal probability. That is, $P(S') = P(S'')$ whenever $|S'| = |S''|$. Let $q_j = P(|\hat{S}| = j)$ for $j = 0, 1, \ldots, n$, DU sampling is uniquely characterized by the vector of probabilities $q$; its density function is given by

$$P(S) = \frac{q_{|S|}}{\binom{n}{|S|}}, \quad S \subseteq [n].$$

- **Nonoverlapping uniform (NU) sampling:** A NU sampling is one which is uniform and which assigns positive probabilities only to sets forming a partition of $[n]$. Let $S^1, S^2, \ldots, S^l$ be a partition of $[n]$, with $|S^j| > 0$ for all $j$. The density function of a NU sampling corresponding to this partition is given by

$$P(S) = \begin{cases} \frac{1}{l} & S \in \{S^1, S^2, \ldots, S^l\} \\ 0, \text{ otherwise} \end{cases}$$

Note that $E[|\hat{S}|] = \frac{n}{l}$. 
Special cases of DU and NU samplings

- **Nice sampling:** Fix $1 \leq \tau \leq n$. A $\tau$-nice sampling is a DU sampling with $q_\tau = 1$.

- **Independent sampling:** Fix $1 \leq \tau \leq n$. A $\tau$-independent sampling is a DU sampling with

$$q_k = \begin{cases} \binom{n}{k} c_k, & k = 1, 2, \ldots, \tau, \\ 0, & k = \tau + 1, \ldots, n, \end{cases}$$

where $c_1 = \left(\frac{1}{n}\right)^\tau$ and $c_k = \left(\frac{k}{n}\right)^\tau - \sum_{i=1}^{k-1} \binom{k}{i} c_i$ for $k \geq 2$.

- **Binomial sampling:** Fix $1 \leq \tau \leq n$ and $0 < p_b \leq 1$. A $(\tau, p_b)$-binomial sampling is defined as a DU sampling with

$$q_k = \binom{\tau}{k} p_b^k (1 - p_b)^{\tau-k}, \quad k = 0, 1, \ldots, \tau.$$ 

Note that $E[|\hat{S}|] = \tau p_b$ and $E[|\hat{S}|^2] = \tau p_b (1 + \tau p_b - p_b)$.

- **Serial sampling:** DU sampling with $q_1 = 1$, NU sampling with $l = n$ and $S^j = \{j\}$ for $j = 1, 2, \ldots, l$.

- **Fully parallel sampling:** DU sampling with $q_n = 1$, NU sampling with $l = 1$ and $S^1 = [n]$.
Let $\emptyset \neq J \subset [n]$ and $\hat{S}$ be any sampling. Further, let $g$ be a block separable function, i.e., $g(x) = \sum_i g_i(x^{(i)})$, then

$$
E[g(x + h\hat{S})] = \sum_{i=1}^n [p_i g_i(x^{(i)} + h^{(i)}) + (1 - p_i) g_i(x^{(i)})].
$$

In addition, if $\hat{S}$ is a uniform sampling, we have

$$
E[g(x + h\hat{S})] = \frac{E[|\hat{S}|]}{n} g(x + h) + \left(1 - \frac{E[|\hat{S}|]}{n}\right) g(x).
$$
Outline

Introduction

Parallel Block Coordinate Descent Methods

Block Sampling

Expected Separable Overapproximation (ESO)

Expected Separable Overapproximation (ESO) of Partially Separable Functions

Iteration Complexity
Expected Separable Overapproximation (ESO)

Definition
Let \( \beta > 0 \), \( w \in \mathbb{R}_+^n \) and let \( \hat{S} \) be a proper uniform sampling. We say that \( f : \mathbb{R}^N \to \mathbb{R} \) admits a \((\beta, w)\)-ESO with respect to \( \hat{S} \) if

\[
E[f(x + h_{[\hat{S}]})] \leq f(x) + \frac{E[|\hat{S}|]}{n} \left( \langle \nabla f(x), h \rangle + \frac{\beta}{2} \|h\|^2_w \right),
\]

holds for all \( x, h \in \mathbb{R}^N \). For simplicity, we write \((f, \hat{S}) \sim \text{ESO}(\beta, w)\). We say that the ESO is monotonic if \( F(x + (h(x))_{[\hat{S}]}) \leq F(x) \) with \( h(x) = \arg \min_{h \in \mathbb{R}^N} H_{\beta, w}(x, h) \) for all \( x \in \text{dom}F \).

- **Inflation.** If \((f, \hat{S}) \sim \text{ESO}(\beta, w)\), then for \( \beta' \geq \beta \) and \( w' \geq w \), \((f, \hat{S}) \sim \text{ESO}(\beta', w')\).

- **Reshuffling.** \((f, \hat{S}) \sim \text{ESO}(c\beta, w) \iff (f, \hat{S}) \sim \text{ESO}(\beta, cw), c > 0\).

- **Strong convexity.** If \((f, \hat{S}) \sim \text{ESO}(\beta, w)\), then \( \beta \geq \mu_f(w) \).
Deterministic Separable Overapproximation (DSO) of Partially Separable Functions

**Theorem**

Assume $f$ is partially separable. Letting $\text{Supp}(h) \equiv \{ i \in [n] : h^{(i)} \neq 0 \}$, for all $h \in \mathbb{R}^N$, we have

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{\max_{J \in \mathcal{J}} |J \cap \text{Supp}(h)|}{2} \|h\|_L^2.$$
Outline

Introduction

Parallel Block Coordinate Descent Methods

Block Sampling

Expected Separable Overapproximation (ESO)

Expected Separable Overapproximation (ESO) of Partially Separable Functions

Iteration Complexity
Uniform samplings

Consider an arbitrary proper sampling $\hat{S}$ and let $\nu = (\nu_1, \ldots, \nu_n)^T$ be defined by

$$\nu_i \equiv \mathbb{E}[\min\{\omega, |\hat{S}|\} | i \in \hat{S}] = \frac{1}{p_i} \sum_{S: i \in S} \mathbb{P}(S) \min\{\omega, |S|\}, \quad i \in [n].$$

**Lemma**
If $\hat{S}$ is proper, then $\mathbb{E}[f(x + h_{\hat{S}})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2} \|h\|^2_{p \odot \nu \odot L}$.

**Theorem**
If $\hat{S}$ is proper and uniform, then

$$(f, \hat{S}) \sim \text{ESO}(1, \nu \odot L).$$

If, in addition, $\mathbb{P}(|\hat{S}| = \tau) = 1$ ($\tau$-uniform), then

$$(f, \hat{S}) \sim \text{ESO}(\min\{\omega, \tau\}, L).$$

Moreover, the ESO is monotonic.
Nonoverlapping uniform samplings

Let $\hat{S}$ be a (proper) nonoverlapping uniform sampling. If $i \in S^j$, for some $j \in \{1, 2, \ldots, l\}$, define

$$\gamma_i \equiv \max_{J \in \mathcal{J}} |J \cap S^j|,$$

and let $\gamma = (\gamma_1, \ldots, \gamma_n)^T$.

**Theorem**

If $\hat{S}$ is a nonoverlapping uniform sampling, then

$$(f, \hat{S}) \sim ESO(1, \gamma \circ L).$$

Moreover, this ESO is monotonic.
Nice samplings

Theorem

If $\hat{S}$ is the $\tau$-nice sampling and $\tau \neq 0$, then

$$(f, \hat{S}) \sim \left(1 + \frac{(\omega - 1)(\tau - 1)}{\max\{1, n - 1\}}, L\right)$$
Theorem

If $\hat{S}$ is a (proper) doubly uniform sampling, then

$$(f, \hat{S}) \sim ESO \left( 1 + \frac{(\omega - 1) \left( \frac{E[|\hat{S}|^2]}{E[|\hat{S}|]} - 1 \right)}{\max\{1, n - 1\}}, L \right)$$
This theorem could have alternatively been proved by writing $\hat{S}$ as a convex combination of nice samplings and applying Theorem 9.

Note that Theorem 15 reduces to that of Theorem 14 in the special case of a nice sampling, and gives the same result as Theorem 13 in the case of the serial and fully parallel samplings.

## Summary of results

In Table 2 we summarize all ESO inequalities established in this section. The first three results cover all the remaining ones as special cases.

<table>
<thead>
<tr>
<th>$\hat{S}$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$w$</th>
<th>Monotonic?</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>$\frac{E[</td>
<td>\hat{S}</td>
<td>]}{n}$</td>
<td>1</td>
</tr>
<tr>
<td>nonoverlapping uniform</td>
<td>$\frac{n}{l}$</td>
<td>1</td>
<td>$\gamma \odot L$</td>
<td>Yes</td>
</tr>
<tr>
<td>doubly uniform</td>
<td>$\frac{E[</td>
<td>\hat{S}</td>
<td>]}{n}$</td>
<td>$1 + \frac{(\omega-1)\left(\frac{E[</td>
</tr>
<tr>
<td>$\tau$-uniform</td>
<td>$\frac{\tau}{n}$</td>
<td>$\min{\omega, \tau}$</td>
<td>$L$</td>
<td>Yes</td>
</tr>
<tr>
<td>$\tau$-nice</td>
<td>$\frac{\tau}{n}$</td>
<td>$1 + \frac{(\omega-1)(\tau-1)}{\max(1,n-1)}$</td>
<td>$L$</td>
<td>No</td>
</tr>
<tr>
<td>$(\tau, p_b)$-binomial</td>
<td>$\frac{\tau p_b}{n}$</td>
<td>$1 + \frac{p_b(\omega-1)(\tau-1)}{\max(1,n-1)}$</td>
<td>$L$</td>
<td>No</td>
</tr>
<tr>
<td>serial</td>
<td>$\frac{1}{n}$</td>
<td>1</td>
<td>$L$</td>
<td>Yes</td>
</tr>
<tr>
<td>fully parallel</td>
<td>1</td>
<td>$\omega$</td>
<td>$L$</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Outline

Introduction

Parallel Block Coordinate Descent Methods

Block Sampling

Expected Separable Overapproximation (ESO)

Expected Separable Overapproximation (ESO) of Partially Separable Functions

Iteration Complexity
Lemma

Fix $x_0 \in \mathbb{R}^N$ and let $\{x_k\}_{k\geq 0}$ be a sequence of random vectors in $\mathbb{R}^N$ with $x_{k+1}$ depending on $x_k$ only. Let $\phi : \mathbb{R}^N \to \mathbb{R}$ be a nonnegative function and define $\xi_k = \phi(x_k)$. Lastly, choose accuracy level $0 < \epsilon < \xi_0$, confidence level $0 < \rho < 1$, and assume that the sequence of random variables $\{\xi_k\}_{k\geq 0}$ is nonincreasing and has one of the following properties:

(i) $\mathbb{E}[\xi_{k+1}|x_k] \leq (1 - \frac{\xi_k}{c_1})\xi_k$, for all $k$, where $c_1 > \epsilon$ is a constant.

(ii) $\mathbb{E}[\xi_{k+1}|x_k] \leq (1 - \frac{1}{c_2})\xi_k$, for all $k$, such that $\epsilon_k \geq \epsilon$, where $c_2 > 1$ is a constant.

If property (i) holds and we choose $K \geq 2 + \frac{c_1}{\epsilon}(1 - \frac{\epsilon}{\xi_0} + \log(\frac{1}{\rho}))$, or if property (ii) holds, and we choose $K \geq c_2 \log(\frac{\xi_0}{\epsilon\rho})$, then

$\mathbb{P}(\xi_K \leq \epsilon) \geq 1 - \rho$.

Let $\xi = F(x_k) - F^*$, we have

$\mathbb{E}[\xi_{k+1}|x_k] \leq (1 - \alpha)\xi_k + \alpha(H_{\beta,w}(x_k, h(x_k)) - F^*)$.

We need a upper bound for $H_{\beta,w}(x_k, h(x_k)) - F^*$. 
Two auxiliary results

From the definition of $h(x)$ and strong convexity of $f$, we have

**Lemma**

For all $x \in \text{dom } F$, $H_{\beta,w}(x, h(x)) \leq \min_{y \in \mathbb{R}^N} \{ F(y) + \frac{\beta - \mu_f(w)}{2} \| y - x \|_w^2 \}$

**Lemma**

(i) Let $x^*$ be an optimal solution, $x \in \text{dom } F$ and let $R = \|x - x^*\|_w$. Then

$$H_{\beta,w}(x, h(x)) - F^* \leq \begin{cases} \left(1 - \frac{F(x) - F^*}{2\beta R^2}\right) (F(x) - F^*), & \text{if } F(x) - F^* \leq \beta R^2 \\ \frac{1}{2} \beta R^2 < \frac{1}{2} (F(x) - F^*), & \text{otherwise} \end{cases}$$

(ii) If $\mu_f(w) + \mu_\Omega(w) > 0$ and $\beta \geq \mu_f(w)$, then for all $x \in \text{dom } F$,

$$H_{\beta,w}(x, h(x)) - F^* \leq \frac{\beta - \mu_f(w)}{\beta + \mu_\Omega(w)} (F(x) - F^*)$$
Theorem

Assume that \((f, \hat{S}) \sim ESO(\beta, w)\), where \(\hat{S}\) is a proper uniform sampling, and 
\[\alpha = \frac{E[\|\hat{S}\|]}{n}.\] Choose \(x_0 \in \mathbb{R}^N\) satisfying 
\[\mathcal{R}_w(x_0, x^*) \equiv \max_x \{\|x - x^*\|_w : F(x) \leq F(x_0)\} < +\infty,\]

where \(x^*\) is an optimal point. Further choose target confidence level \(0 < \rho < 1\), target accuracy level \(\epsilon > 0\) and iteration counter \(K\) in any of the following two ways:

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

\[K \geq 2 + \frac{2(\frac{\beta}{\alpha}) \max \left\{ \mathcal{R}_w^2(x_0, x^*), \frac{F(x_0) - F^*}{\beta} \right\}}{\epsilon} \left(1 - \frac{\epsilon}{F(x_0) - F^*} + \log \left(\frac{1}{\rho}\right)\right)\]

(ii) \(\epsilon < \min \{2(\frac{\beta}{\alpha}) \mathcal{R}_w^2(x_0, x^*), F(x_0) - F^*\}\) and

\[K \geq 2\frac{\beta}{\alpha} \max \left\{ \frac{\mathcal{R}_w^2(x_0, x^*)}{\epsilon}, \frac{1}{\beta} \right\} \log \left(\frac{F(x_0) - F^*}{\epsilon \rho}\right)\]

If \(\{x_k\}, k \geq 0\) are the random iterates of PCDM2 or PCDM1 if the ESO is monotonic, then \(\mathbb{P}(F(x_K) - F^* \leq \epsilon) \geq 1 - \rho.\)
Since $F(x_k) \leq F(x_0)$ for all $k$, we have $\|x_k - x^*\|_w \leq \mathcal{R}_w(x_0, x^*)$.

$$\mathbb{E}[\xi_{k+1}|x_k] \leq (1 - \alpha)\xi_k + \alpha(H_{\beta,w}(x_k, h(x_k)) - F^*)$$

$$\leq (1 - \alpha)\xi_k + \alpha \max\left\{1 - \frac{\xi_k}{2\beta\|x_k - x^*\|_w^2}, \frac{1}{2}\right\} \xi_k$$

$$= \max\left\{1 - \frac{\alpha \xi_k}{2\beta\|x_k - x^*\|_w^2}, 1 - \frac{\alpha}{2}\right\} \xi_k$$

$$\leq \max\left\{1 - \frac{\alpha \xi_k}{2\beta \mathcal{R}_w^2(x_0, x^*)}, 1 - \frac{\alpha}{2}\right\} \xi_k$$

- Let $c_1 = 2\frac{\beta}{\alpha} \max\{\mathcal{R}_w^2(x_0, x^*), \frac{\xi_0}{\beta}\}$, we have $\mathbb{E}[\xi_{k+1}|x_k] \leq (1 - \frac{\xi_k}{c_1})\xi_k$

- Let $c_2 = 2\frac{\beta}{\alpha} \max\{\frac{\mathcal{R}_w^2(x_0, x^*)}{e}, \frac{1}{\beta}\}$, we have $\mathbb{E}[\xi_{k+1}|x_k] \leq (1 - \frac{1}{c_2})\xi_k$. 
Speedup factor for convex case

The iteration complexity of our methods in the convex case is $O\left(\frac{\beta}{\alpha} \frac{1}{\epsilon}\right)$. Let us define the parallelization speedup factor by

$$\text{parallelization speedup factor} = \frac{\frac{\beta}{\alpha} \text{ of the serial method}}{\frac{\beta}{\alpha} \text{ of a parallel method}} \frac{n}{\frac{\beta}{\alpha} \text{ of a parallel method}}$$
## Speedup for DU samplings

The important message of the above theorem is that the iteration complexity of our methods in the convex case is $O(\beta^\alpha \epsilon)$. Note that for the serial method (PCDM1 used with $\hat{S}$ being the serial sampling) we have $\alpha = 1/n$ and $\beta = 1/(\hat{S}, \hat{S})$, and hence $\beta^\alpha = n$. It will be interesting to study the parallelization speedup factor defined by

$$\frac{\mathbb{E}[||\hat{S}||]}{(\omega - 1)(\mathbb{E}[||\hat{S}||^2]/\mathbb{E}[||\hat{S}||]) - 1)}$$

$$1 + \frac{\max(1, n-1)}{(\omega - 1)(\mathbb{E}[||\hat{S}||^2]/\mathbb{E}[||\hat{S}||]) - 1)}$$

Table 3, computed from the data in Table 2, gives expressions for the parallelization speedup factors for PCDM based on a DU sampling (expressions for 4 special cases are given as well).

### Table 3: Convex $F$: Parallelization speedup factors for DU samplings. The factors below the line are special cases of the general expression. Maximum speedup is naturally obtained by the fully parallel sampling: $n/\omega$.

<table>
<thead>
<tr>
<th>$\hat{S}$</th>
<th>Parallelization speedup factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>doubly uniform</td>
<td>$\frac{\mathbb{E}[</td>
</tr>
<tr>
<td>($\tau, p_b$)-binomial</td>
<td>$\frac{\mathbb{E}[</td>
</tr>
<tr>
<td>$\tau$-nice</td>
<td>$\frac{\tau}{1 + \frac{(\omega - 1)(\tau - 1)}{\max(1, n-1)}}$</td>
</tr>
<tr>
<td>fully parallel</td>
<td>$\frac{n}{\omega}$</td>
</tr>
<tr>
<td>serial</td>
<td>$1$</td>
</tr>
</tbody>
</table>
Iteration complexity: strongly convex case

\( F(x_k) \) converges to \( F^* \) linearly, with high probability.

**Theorem**

Assume \( F \) is strongly convex with \( \mu_f(w) + \mu_\Omega(w) > 0 \). Further, assume \((f, \hat{S}) = ESO(\beta, w)\), where \( \hat{S} \) is a proper uniform sampling and let

\[
\alpha = \frac{E[|\hat{S}|]}{n}.
\]

Choose initial point \( x_0 \in \mathbb{R}^N \), target confidence level \( 0 < \rho < 1 \), target accuracy level \( 0 < \epsilon < F(x_0) - F^* \), and

\[
K \geq \frac{1}{\alpha \mu_f(w) + \mu_\Omega(w)} \log \left( \frac{F(x_0) - F^*}{\epsilon \rho} \right).
\]

If \( \{x_k\} \) are the random points generated by PCDM1 or PCDM2, then

\[
P(F(x_K) - F^* \leq \epsilon) \geq 1 - \rho.
\]

\[
E[\xi_{k+1}|x_k] \leq (1 - \alpha)\xi_k + \alpha(H_{\beta,w}(x_k, h(x_k)) - F^*)
\]

\[
\leq \left(1 - \alpha \frac{\mu_f(w) + \mu_\Omega(w)}{\beta + \mu_\Omega(w)}\right) \xi_k \equiv (1 - \gamma)\xi_k
\]

\[
P[\xi_K > \epsilon] \leq \frac{E[\xi_k]}{\epsilon} \leq \frac{(1 - \gamma)^K \xi_0}{\epsilon} \leq \rho
\]
Strongly convex case

- Without the requirement of monotonicity.

parallelization speedup factor = \[\frac{n}{\mu_f(w) + \mu_\Omega(w)} \cdot \frac{\beta + \mu_\Omega(w)}{\alpha (1 + \mu_\Omega(w))}\] of a parallel method

- The speedup factor is independent of \(\mu_f(w)\).
- When \(\mu_\Omega(w) \to 0\), the speedup factor approaches the factor in the non-strongly convex case.
- For large values of \(\mu_\Omega(w)\), the speedup factor is approximately equal \(\alpha n = E[|\hat{S}|]\), which is the average number of blocks updated in a single parallel iteration.
- Speedup factor does not depend on the degree of partial separability of \(f\).