# On optimal Probabilities in Stochastic Coordinate Descent Methods — 'NSync

Peter Richtárik Martin Takáč

*University of Edinburgh*

NAIS Centre for Numerical Algorithms and Intelligent Software

Simons Institute for the Theory of Computing

Mathematics for Vast Digital Resources

## 1. The Problem

We are interested in solving the optimization problem

$$\min_{x\in\mathbb{R}^n} \phi(x) \qquad \text{(P)}$$

where $\phi(x)$ is differentiable and $\gamma$-**strongly convex** wrt a weighted Euclidean norm with weights $v_1,\dots,v_n>0$. That is, for all $x,h\in\mathbb{R}^n$,

$$\phi(x+h)\ge \phi(x)+\sum_{i=1}^n \nabla_i\phi(x)h_i+\frac{\gamma}{2}\sum_{i=1}^n v_i h_i^2, \qquad (1)$$

where $\nabla_i\phi(x)=e_i^T\nabla\phi(x)$ and $e_i$ is the $i$th unit coordinate vector.

## 2. The Algorithm ('NSync)

**Pre-processing:**
1. Assign a probability $p_S\ge 0$ to every subset $S$ of $\{1,\dots,n\}$ such that $\sum_S p_S=1$
2. For $i=1,2,\dots,n$ we pick a stepsize $w_i>0$

**Algorithm ('NSync)**

**Input:** Initial point $x_0\in\mathbb{R}^n$
**for** $k=0,1,2,\dots$ **do**
1. Choose **random** $\hat{S}\subseteq\{1,2,\dots,n\}$ such that $\mathbf{Prob}(\hat{S}=S)=p_S$
2. $x_{k+1}\leftarrow x_k-\sum_{i\in\hat{S}}\frac{1}{w_i}\nabla_i\phi(x_k)e_i$

## 3. Nonuniform ESO

**Assumption:** We assume that $\phi,\hat{S}$ admit the following nonuniform expected separable overapproximation: For all $x,h\in\mathbb{R}^n$

$$\mathbf{E}\left[\phi\left(x+\sum_{i\in\hat{S}}h_ie_i\right)\right]\le\phi(x)+\sum_{i=1}^n p_i(\nabla\phi(x))_ih_i+\frac{1}{2}\sum_{i=1}^n p_iw_ih_i^2. \qquad (2)$$

where $p_i=\mathbf{Prob}(i\in\hat{S})=\sum_{S:i\in S}p_S$.

## 4. Convergence Theorem

**Theorem 1: Iteration Complexity Guarantees**

*Let (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*

$$\Lambda:=\max_i\frac{w_i}{p_iv_i}. \qquad (3)$$

*If* $\{x_k\}$ *are the random iterates generated by 'NSync, then*

$$K\ge\frac{\Lambda}{\gamma}\log\left(\frac{\phi(x_0)-\phi^*}{\epsilon\rho}\right)\quad\Longrightarrow\quad\mathbf{Prob}(\phi(x_K)-\phi^*\le\epsilon)\ge 1-\rho. \qquad (4)$$

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

## 5. Application

Consider problem (P) with $\phi$ of the form

$$\phi(x):=f(x)+\frac{\gamma}{2}\sum_{i=1}^n v_ix_i^2. \qquad (5)$$

**Smoothness Assumption:** $f$ has Lipschitz gradient wrt the coordinates. That is, for some $L_i>0$ and all $x\in\mathbb{R}^n$ and $t\in\mathbb{R}$: $|\nabla_if(x)-\nabla_if(x+te_i)|\le L_i|t|$.

**Partial Separability Assumption:** $f(x)=\sum_{J\in\mathcal{J}}f_J(x)$, where $f_J$ are differentiable convex functions such that $f_J$ depends on coordinates $i\in J\subseteq\{1,2,\dots,n\}$ only. Let $\omega:=\max_J|J|$. We say that $f$ is *separable of degree* $\omega$.

**Nonuniform sampling $\hat{S}$.** Fix $\tau\in\{1,\dots,n\}$ and $c\ge1$ and let $S_1,\dots,S_c$ be a collection of (possibly overlapping) subsets of $\{1,\dots,n\}$ such that $|S_j|\ge\tau$ for all $i$ and $\cup_{j=1}^cS_j=\{1,\dots,n\}$. Moreover, let $q=(q_1,\dots,q_c)>0$ be a probability vector. Now, $\hat{S}$ is generated as follows:

1. Pick $j\in\{1,\dots,c\}$ with probability $q_j$,
2. Draw $\hat{S}_j\subseteq S_j$ with cardinality $\tau$, uniformly at random.

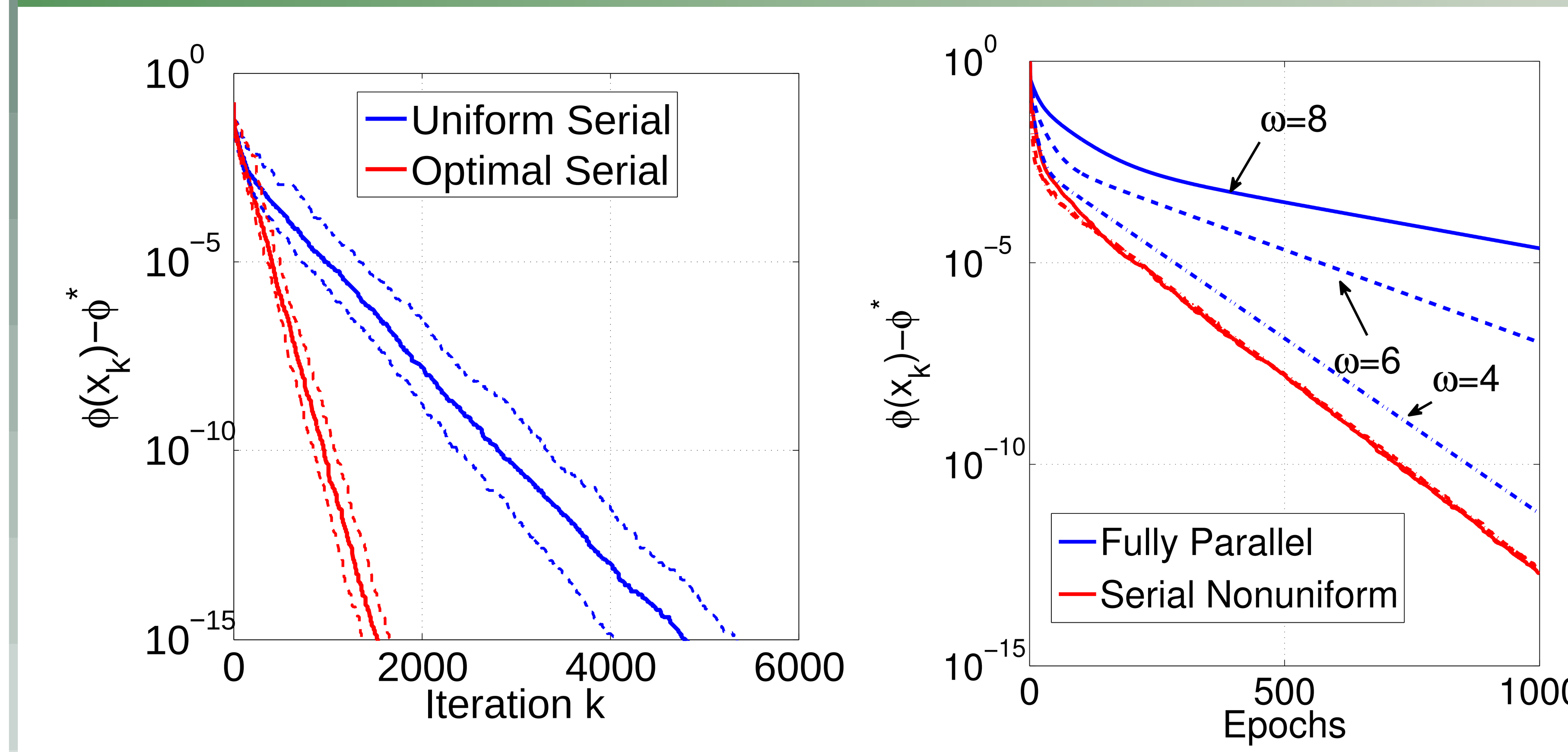
**Theorem 2: ESO Parameters**

*Under assumptions mentioned above, (2) holds with*

$$w_i\ge w_i^*:=\tfrac{L_i+v_i}{p_i}\sum_{j=1}^c q_j\tfrac{\tau}{|S_j|}\delta_{ij}\left(1+\tfrac{(\tau-1)(\omega_j-1)}{\max\{1,|S_j|-1\}}\right),\quad i\in\{1,\dots,n\}, \qquad (6)$$

*where* $\omega_j:=\max_{J\in\mathcal{J}}|J\cap S_j|\le\omega$, *and* $\delta_{ij}=\begin{cases}1, & \text{if } i\in S_j,\\ 0, & \text{otehwise.}\end{cases}$

## 7. Experiments




**LEFT:** $A\in\mathbb{R}^{2\times30}$, $\gamma=1$, $v_1=0.05$, $v_i=1$ for $i\ne1$ and $L_i=1$ for all $i$. We compare the US method ($p_i=1/n$, blue) with the OS method ($p_i$ are optimal, red). Dashed lines = 95% confidence intervals (line in the middle is the average).

**RIGHT:** Nonuniform serial (NS) method can be faster than the fully parallel (FP) variant ($m=8$, $n=10$).

## 6. Optimal Probabilities

We can *design* optimal probabilities using (6) for a sampling (characterized by the sets $S_j$ and probabilities $q_j$) that *minimizes* $\Lambda$, which in view of (4) *optimizes the convergence rate* of the method.

**Serial setting.** Let $c=n$, with $S_i=\{i\}$, $\mathbf{Prob}(|\hat{S}|=1)=1$ and $p_i=q_i$ for all $i\in\{1,\dots,n\}$. From (6) we get $w_i=w_i^*=L_i+v_i$.
Minimizing $\Lambda$ in (3) over the probability vector $p$ gives the *optimal probabilities* (we refer to this as the *optimal serial (OS)* method) and *optimal complexity*

$$p_i^*=\frac{(L_i+v_i)/v_i}{\sum_j(L_j+v_j)/v_j},\quad \Lambda_{OS}=n+\sum_i\frac{L_i}{v_i}.$$

Note that the *uniform sampling*, $p_i=1/n$ for all $i$, leads to $\Lambda_{US}:=n+n\max_jL_j/v_j$ (we call this the *uniform serial (US)* method), which can be much larger than $\Lambda_{OS}$.

**Fully Parallel (FP) setting.** Set $c=1$ and $\tau=n$, which yields $\Lambda_{FP}=\omega+\omega\max_jL_j/v_j$. Since $\omega\le n$, it is clear that $\Lambda_{US}\ge\Lambda_{FP}$. However, for large enough $\omega$, we have $\Lambda_{FP}\ge\Lambda_{OS}$.

**The optimal serial method can be faster than the fully parallel method!**

**Parallel setting.** Fix $\tau$ and sets $S_j$, $j=1,2,\dots,c$, and define $\theta:=\max_j\left(1+\frac{(\tau-1)(\omega_j-1)}{\max\{1,|S_j|-1\}}\right)$. Consider running 'NSync with stepsizes $w_i=\theta(L_i+v_i)$.
The complexity of 'NSync is determined by

$$\Lambda=\max_i\tfrac{w_i}{p_iv_i}=\tfrac{\theta}{\tau}\max_i\left(1+\tfrac{L_i}{v_i}\right)\left(\sum_{j=1}^cq_j\tfrac{\delta_{ij}}{|S_j|}\right)^{-1}.$$

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

## 8. References


[1] Richtárik, P. and Takáč, M.: On optimal probabilities in stochastic coordinate descent methods, 2013

[2] Richtárik, P., Takáč, M.: Parallel coordinate descent methods for big data optimization, 2012

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