Fast Distributed Coordinate Descent for Non-Strongly Convex Losses

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

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.
FAST DISTRIBUTED COORDINATE DESCENT FOR NON-STRONGLY CONVEX LOSSES*

Olivier Fercoq  Zheng Qu  Peter Richtárik  Martin Takáč

School of Mathematics, University of Edinburgh, Edinburgh, EH9 3JZ, United Kingdom

ABSTRACT
We propose an efficient distributed randomized coordinate descent method for minimizing regularized non-strongly convex loss functions. The method attains the optimal $O(1/k^2)$ convergence rate, where $k$ is the iteration counter. The core of the method is the theoretical study of stepsize parameters. We have implemented the method on Archer—the largest supercomputer in the UK—and show that the method is capable of solving a (synthetic) LASSO optimization problem with 50 billion variables.

Index Terms— Coordinate descent, distributed algorithms, acceleration.

1. INTRODUCTION
In this paper we are concerned with solving regularized convex optimization problems in huge dimensions in cases when the loss being minimized is not strongly convex. That is, we consider the problem

$$
\min_{x \in \mathbb{R}^d} L(x) := f(x) + R(x),
$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex differentiable loss function, $d$ is huge, and $R(x) \equiv \sum_{i=1}^n R_i(x^i)$, where $R_i : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ are (possibly nonsmooth) convex regularizers and $x^i$ denotes the $i$-th coordinate of $x$. We make the following technical assumption about $f$: there exists a $n$-by-$d$ matrix $A$ such that for all $x, h \in \mathbb{R}^d$,

$$
f(x + h) \leq f(x) + (f'(x))^\top h + \frac{1}{2} h^\top A^\top A h.
$$

For examples of regularizers $R$ and loss functions $f$ satisfying the above assumptions, relevant to machine learning applications, we refer the reader to [1][2][3].

It is increasingly the case in modern applications that the data describing the problem (encoded in $A$ and $R$ in the above model) is so large that it does not fit into the RAM of a single computer. In such a case, unless the application at hand can tolerate slow performance due to frequent HDD reads/writes, it is often necessary to distribute the data among the nodes of a cluster and solve the problem in a distributed manner.

While efficient distributed methods exist in cases when the regularized loss $L$ is strongly convex (e.g., Hydra [3]), here we assume that $L$ is not strongly convex. Problems of this type arise frequently: for instance, in the SVM dual, $f$ is typically a non-strongly convex quadratic, $d$ is the number of examples, $n$ is the number of features, $A$ encodes the data, and $R$ is a $0$-$\infty$ indicator function encoding box constraints (e.g., see [2]). If $d > n$, then $L$ will not be strongly convex.

In this paper we propose “Hydra$^2$” (Hydra “squared”; Algorithm [1]) — the first distributed stochastic coordinate descent (CD) method with fast $O(1/k^2)$ convergence rate for our problem. The method can be seen both as a specialization of the APPROX algorithm [4] to the distributed setting, or as an accelerated version of the Hydra algorithm [3] (Hydra converges as $O(1/k)$ for our problem). The core of the paper forms the development of new stepsizes, and new efficiently computable bounds on the stepsizes proposed for distributed CD methods in [3]. We show that Hydra$^2$ is able to solve a big data problem with $d$ equal to 50 billion.

2. THE ALGORITHM
Assume we have $c$ nodes/computers available. In Hydra$^2$ (Algorithm [1]), the coordinates $i \in [d] := \{1, 2, \ldots, d\}$ are first partitioned into $c$ sets $\{P_l, l = 1, \ldots, c\}$, each of size $|P_l| = s := d/c$. The columns of $A$ are partitioned accordingly, with those belonging to partition $P_l$ stored on node $l$. During one iteration, all computers $l \in \{1, \ldots, c\}$, in parallel, pick a subset $S_l$ of $\tau$ coordinates from those they own, i.e., from $P_l$, uniformly at random, where $1 \leq \tau \leq s$ is a parameter of the method (Step 6). From now on we denote by $\hat{S}$ the union of all these sets, $\hat{S} := \cup_l S_l$, and refer to it by the name distributed sampling.

Hydra$^2$ maintains two sequences of iterates: $u_k, z_k \in \mathbb{R}^d$. Note that this is usually the case with accelerated/fast gradient-type algorithms [5][6][7]. Also note that the output of the method is a linear combination of these two vectors. These iterates are stored and updated in a distributed way, with the $i$-th coordinate stored on computer $l$ if $i \in P_l$. Once computer $l$ picks $\hat{S}_l$, it computes for each $i \in \hat{S}_l$ an update scalar $t_k^i$, which is then used to update $z_k^i$ and $u_k^i$, in parallel (using the multicore capability of computer $l$).

The main work is done in Step [8], where the update scalars

*THANKS TO EPSRC FOR FUNDING VIA GRANTS EP/K02325X/1, EP/1017127/1 AND EP/G036136/1.
Algorithm 1 Hydra

1. **INPUT:** \( \{P_i\}_{i=1}^c, l \leq \tau \leq s, \{D_{ii}\}_{i=1}^d, z_0 \in \mathbb{R}^d \)
2. set \( \theta_0 = \tau / s \) and \( u_0 = 0 \)
3. for \( k \geq 0 \) do
4. \( z_{k+1} \leftarrow z_k, u_{k+1} \leftarrow u_k \)
5. for each process \( i \in \{1, \ldots, c\} \) in parallel do
6. \( \tilde{t}_k^i \leftarrow \arg\min_{t_k} f'_i(\theta_k^i u_k + z_k) t + \frac{\|D_{ii}\|_2}{2} t^2 + \mathcal{R}_e(z_k^i + t) \)
7. \( z_{k+1}^i \leftarrow z_k^i + \tilde{t}_k^i, u_{k+1}^i \leftarrow u_k^i - (\frac{1}{\theta_k^i} - \frac{1}{\tau \theta_k^i}) \tilde{t}_k^i \)
8. end parallel for
9. end parallel for
10. \( \theta_{k+1} = \frac{1}{2}(\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2) \)
11. end for
12. **OUTPUT:** \( \theta_k^2 u_k + z_{k+1} \)

The partial derivatives are computed at \( (\theta_k^2 u_k + z_k) \). For the algorithm to be efficient, the computation should be performed without computing the sum \( \theta_k^2 u_k + z_k \). As shown in [4], this can be done for functions \( f \) of the form

\[ f(x) = \sum_{j=1}^m \phi_j (e_j^T A x), \]

where for all \( j \), \( \phi_j \) is a one-dimensional differentiable (convex) function. Indeed, let \( D_i \) be the set of such \( j \) for which \( A_{ji} \neq 0 \). Assume we store and update \( r_{uk} = A u_k \) and \( r_{z} = A z_k \), then

\[ \nabla_i f(\theta_k^2 u_k + z_k) = \sum_{j \in D_i} A_{ji}^T \phi'_j(\theta_k^2 r_{uk} + r_{z}). \]

The average cost for this computation is \( \frac{1}{\tau} \sum_{i=1}^c \mathcal{O}(\|D_i\|) \), i.e., each processor needs to access the elements of only one column of the matrix \( A \).

Steps 8 and 9 depend on a deterministic scalar sequence \( \theta_k \), which is being updated in Step 12 as in [6]. Note that by taking \( \theta_k = \theta_0 \) for all \( k \), \( u_k \) remains equal to 0, and Hydra reduces to Hydra.

The output of the algorithm is \( x_{k+1} = (\theta_k^2 u_{k+1} + z_{k+1}) \). We only need to compute this vector sum at the end of the execution and when we want to track \( L(x_k) \). Note that one should not evaluate \( x_k \) and \( L(x_k) \) at each iteration since these computations have a non-negligible cost.

3. CONVERGENCE RATE

The magnitude of \( D_{ii} > 0 \) directly influences the size of the update \( \tilde{t}_k^i \). In particular, note that when there is no regularizer \( (R_i = 0) \), then \( \tilde{t}_k^i = 2\tau f'_i(\theta_k^2 u_k + z_k)(s_0 D_{ii})^{-1} \).

That is, small \( D_{ii} \) leads to a larger “step” \( \tilde{t}_k^i \) which is used to update \( z_k^i \) and \( u_k^i \). For this reason we refer to \( \{D_{ii}\}_{i=1}^d \) as stepsize parameters. Naturally, some technical assumptions on \( \{D_{ii}\}_{i=1}^d \) should be made in order to guarantee the convergence of the algorithm. The so-called ESO (Expected Separable Overapproximation) assumption has been introduced in this scope. For \( h \in \mathbb{R}^d \), denote \( h^S := \sum_{i \in S} h^t e_i \), where \( e_i \) is the \( i \)th unit coordinate vector.

**Assumption 3.1 (ESO).** Assume that for all \( x \in \mathbb{R}^d \) and \( h \in \mathbb{R}^d \) we have

\[ E[f(x + h^S)] \leq f(x) + \frac{E[|S|]}{d}((f'(x))^T h + \frac{1}{2} h^T D h), \]

where \( D \) is a diagonal matrix with diag. elements \( D_{ii} > 0 \) and \( S \) is the distributed sampling described above.

The above ESO assumption involves the smooth function \( f \), the sampling \( S \) and the parameters \( \{D_{ii}\}_{i=1}^d \). It has been first introduced by Richtárik and Takáč [2] for proposing a generic approach in the convergence analysis of the Parallel Coordinate Descent Methods (PCDM). Their generic approach boils down the convergence analysis of the whole class of PCDMs to the problem of finding proper parameters which make the ESO assumption hold. The same idea has been extended to the analysis of many variants of PCDM, including the Accelerated Coordinate Descent algorithm [3] (APPROX) and the Distributed Coordinate Descent method [3] (Hydra).

In particular, the following complexity result, under the ESO assumption 3.1, can be deduced from [4, Theorem 3] using Markov inequality.

**Theorem 3.2 ([4]).** Let \( x_s \in \mathbb{R}^d \) be any optimal point of \( \{1\} \),
\[
\text{for} \quad 0 < \rho < 1 \quad \text{and} \quad 0 < \epsilon < L_0 \quad \text{and let}
\[
C_1 := (1 - \frac{\epsilon}{L_0}), \quad C_2 := \frac{1}{2} d^2_0 D_0 d_0.
\]
Choose \( 0 < \rho < 1 \) and \( 0 < \epsilon < L_0 \) and
\[
k \geq \frac{2 \sqrt{s}}{\tau} \left( \sqrt{\frac{C_1 + C_2}{\rho \epsilon}} - 1 \right) + 1. \quad (4)
\]

Then under Assumption 3.1, the iterates \( \{x_k\}_{k=1}^\infty \) of Algorithm 3 satisfy
\[
\text{Prob}(L(x_k) - L(x_s) \leq \epsilon) \geq 1 - \rho.
\]

The bound on the number of iterations \( k \) in [4] suggests to choose the smallest possible \( \{D_{ii}\}_{i=1}^d \) satisfying the ESO assumption 3.1.

4. NEW STEPSIZES

In this section we propose new stepsize parameters \( D_{ii} \) for which the ESO assumption is satisfied.

For any matrix \( G \in \mathbb{R}^{d \times d} \), denote by \( D^G \) the diagonal matrix such that \( D^G_{ii} = G_{ii} \) for all \( i \) and \( D^G_{ij} = 0 \) for \( i \neq j \). We denote by \( B^G \in \mathbb{R}^{d \times d} \) the block matrix associated with the partition \( \{P_1, \ldots, P_c\} \) such that \( B^G_{ij} = G_{ij} \) whenever \( \{i, j\} \subset P_l \) for some \( l \), and \( B^G_{ij} = 0 \) otherwise.
Let \( \omega_j \) be the number of “partitions active at row \( j \)”, i.e., the number of indexes \( l \in \{1, \ldots, c\} \) for which the set \( \{i \in P_l : A_{li} \neq 0\} \) is nonempty. Note that since \( A \) does not have an empty row or column, we know that \( 1 \leq \omega_j \leq n \) and \( 1 \leq \omega_j' \leq c \). Moreover, we have the following characterization of these two quantities. For notational convenience we denote \( M_j = A^\top_j A_j \) for all \( j \in \{1, \ldots, n\} \).

**Lemma 4.1.** For \( j \in \{1, \ldots, n\} \) we have:

\[
\omega_j = \max\{x^\top M_j x : x^\top D_j M_j x \leq 1\},
\]

\[
\omega_j' = \max\{x^\top M_j x : x^\top B_j M_j x \leq 1\}.
\]

*Proof.* For \( l \in \{1, \ldots, c\} \) and \( y \in \mathbb{R}^d \), denote \( y^{(l)} := (y')_{i \in P_l}. \) That is, \( y^{(l)} \) is the subvector of \( y \) composed of coordinates belonging to partition \( P_l. \) Then we have:

\[
x^\top M_j x = \sum_{i=1}^c (A^{(l)}_j x^{(l)})^2, \quad x^\top B_j M_j x = \sum_{i=1}^c (A^{(l)}_j x^{(l)})^2.
\]

Let \( S' = \{l : A^{(l)}_j \neq 0\} \), then \( \omega_j' = |S'| \) and by Cauchy-Schwarz we have

\[
(\sum_{l \in S'} A^{(l)}_j x^{(l)})^2 \leq \omega_j' \sum_{l \in S'} (A^{(l)}_j x^{(l)})^2.
\]

Equality is reached when \( A^{(l)}_j x^{(l)} = \alpha \) for some constant \( \alpha \) for all \( l \in S' \) (this is feasible since the subsets \( \{P_1, \ldots, P_c\} \) are disjoint). Hence, we proved \( \omega_j' \). The characterization \( \omega_j \) follows from \( \omega_j' \) by setting \( c = d \).

We shall also need the following lemma.

**Lemma 4.2 (2).** Fix arbitrary \( G \in \mathbb{R}^{d \times d} \) and \( x \in \mathbb{R}^d \) and let \( s_1 = \max\{1, s - 1\} \). Then \( E[(x^\top S G x^\top S)] \) is equal to

\[
\frac{1}{s} [\alpha_1 x^\top D G x + \alpha_2 x^\top G x + \alpha_3 x^\top (G - B) x],
\]

where \( \alpha_1 = 1 - \frac{s - 1}{s_1}, \alpha_2 = \frac{s - 1}{s_1}, \alpha_3 = \frac{s}{s} - \frac{s - 1}{s_1}. \)

Below is the main result of this section.

**Theorem 4.1.** For a convex differentiable function \( f \) satisfying \( (2) \) and a distributed sampling \( S \) described in Section 2, the ESO assumption \( (3) \) is satisfied for

\[
D^2_{ii} \equiv D_{ii} = \beta^* \sum_{j=1}^n A^2_{ij}, \quad i = 1, \ldots, d \quad (13)
\]

where

\[
\beta^* := \beta^*_1 + \beta^*_2 \quad (14)
\]

5. EXISTING STEPSIZES AND NEW BOUNDS

We now describe stepsizes previously suggested in the literature. For simplicity, let \( M := A^\top A. \) Define:

\[
\sigma := \max\{x^\top M x : x \in \mathbb{R}^d, x^\top D M x \leq 1\},
\]

\[
\sigma' := \max\{x^\top M x : x \in \mathbb{R}^d, x^\top B M x \leq 1\}.
\]

The quantities \( \sigma \) and \( \sigma' \) are identical to those defined in \( (3) \) (although the definitions are slightly different). The following stepsize parameters have been introduced in \( (3) \):

**Lemma 5.1 (3).** The ESO assumption \( (3) \) is satisfied for

\[
D^2_{ii} \equiv D_{ii} = \beta^* \sum_{j=1}^n A^2_{ij}, \quad i = 1, \ldots, d \quad (13)
\]

where \( \beta^* := \beta^*_1 + \beta^*_2 \) and

\[
\beta^*_1 = 1 + \frac{(s - 1)(s - 1)}{s_1}, \quad \beta^*_2 = \frac{s}{s_1} - \frac{s - 1}{s_1} \quad (14)
\]

5.1. Known bounds on existing stepsizes

In general, the computation of \( \sigma \) and \( \sigma' \) can be done using the Power iteration method \( (9) \). However, the number of operations required in each iteration of the Power method is at least twice as the number of nonzero elements in \( A \). Hence the only computation of the parameters \( \sigma \) and \( \sigma' \) would already require quite a few number of passes through the data. Instead, if one could provide some easily computable upper bound for \( \sigma \) and \( \sigma' \), where by ‘easily computable’ we mean computable by only one pass through the data, then we can run Algorithm \( (11) \) immediately without spending too much time on the computation of \( \sigma \) and \( \sigma' \). Note that the ESO assumption will still hold when \( \sigma \) and \( \sigma' \) in \( (13) \) are replaced by some upper bounds.
In [3], the authors established the following bound

\[ \beta^* \leq 2\beta_1^* = 2\left(1 + \frac{(\tau-1)(\tau-1)}{s-1}\right), \]  

(15)

which is independent of the partition \( \{P_i\}_{i=1,...,c} \). This bound holds for \( \tau \geq 2 \). Further, they showed that

\[ \sigma \leq \max_j \omega_j. \]  

(16)

Then in view of (15), (16) and Lemma 5.3, the ESO assumption [3, Lemma 2] is also satisfied for the following easily computable parameters:

\[ D_{ii}^3 \equiv D_{ii} = 2\left(1 + \frac{\tau-1}{s-1}(\max_j \omega_j - 1)\right)\sum_{j=1}^n A_{ji}^2. \]  

(17)

### 5.2. Improved bounds on existing stepsizes

In what follows, we show that both (15) and (16) can be improved so that smaller parameters \( \{D_{ii}\}_{i=1}^d \) are allowed in the algorithm.

**Lemma 5.2.** Suppose \( \tau \geq 2 \) (note that then \( s \geq \tau \geq 2 \)). For all \( 1 \leq \eta \leq s \) the following holds

\[ \left(\frac{s}{\tau} - \frac{\tau-1}{s-1}\right)\eta \leq \frac{1}{\tau-1}\left(1 + \frac{(\tau-1)(\tau-1)}{s-1}\right) \]  

(18)

**Proof.** Both sides of the inequality are linear functions of \( \eta \). It therefore suffices to verify that the inequality holds for \( \eta = 1 \) and \( \eta = s \), which can be done.

The following result is an improvement on [3, Lemma 2].

**Lemma 5.3.** If \( \tau \geq 2 \), then \( \beta^* \leq (1 + \frac{1}{\tau-1})\beta_1^* \).

**Proof.** We only need to apply Lemma 5.2 to (14) with \( \eta = \sigma \), and additionally use the bound \( \frac{\tau^2 - 1}{\tau-1} \leq 1 \). This gives \( \beta_1^* \leq \frac{\sigma^2}{\sigma_2} \leq \beta_1^*/\tau - 1 \), from which the result follows.

The above lemma leads to an important insight: as long as \( \tau \geq 2 \), the effect of partitioning the data (across the nodes) on the iteration complexity of Hydra is negligible, and vanishes as \( \tau \) increases.

Indeed, consider the complexity bound provided by stepsizes \( D_2 \) given in [3], used within Theorem 3.2 and note that the effect of the partitioning on the complexity is through \( \sigma' \) only (which is not easily computable!), which appears in \( \beta_2^* \) only. However, by the above lemma, \( \beta_1^* \leq \beta^* \leq \beta_2^* \) for any partitioning (to \( c \) sets each containing \( s \) coordinates). Hence, i) each partitioning is near-optimal and ii) one does not even need to know \( \sigma' \) to run the method as it is enough to use stepsizes \( D_2^* \) with \( \beta^* \) replaced by the upper bound provided by Lemma 5.3.

Note that the same reasoning can be applied to the Hydra method [3] since the same stepsizes can be used there. Also note that for \( \tau = 1 \) the effect of partitioning can have a dramatic effect on the complexity.

**Remark 5.1.** Applying the same reasoning to the new stepsizes \( \alpha_i^* \) we can show that \( \alpha_i^* \leq (1 + 1/(\tau-1))\alpha_i^* \) for all \( j \in \{1,...,n\} \). This is not as useful as Lemma 5.2 since \( \alpha_i^* \) does not need to be approximated (\( \omega'_j \) is easily computable).

We next give a tighter upper bound on \( \sigma \) than (16).

**Lemma 5.4.** If we let

\[ v_i := \frac{\sum_{j=1}^n \omega_j A_{ji}^2}{\sum_{j=1}^n A_{ji}^2}, \quad i = 1,\ldots,d, \]  

(19)

then \( \sigma \leq \tilde{\sigma} := \max_i v_i \).

**Proof.** In view of Lemma 4.1 we know that

\[ M = \sum_{j=1}^n M_j \preceq \sum_{j=1}^n \omega_j D_{Mj} = \text{Diag}(v)D^M \preceq \max_i v_i D^M. \]

The rest follows from the definition of \( \sigma \).

By combining Lemma 5.2, Lemma 5.3 and Lemma 5.4, we obtain the following smaller (compared to (17)) admissible and easily computable stepsize parameters:

\[ D_{ii}^4 \equiv D_{ii} = \frac{\tau}{\tau-1}\left(1 + \frac{(\bar{\sigma} - 1)(\tau-1)}{s-1}\right)\sum_{j=1}^n A_{ji}^2. \]  

(20)

### 6. COMPARISON OF OLD AND NEW STEPSIZES

So far we have seen four different admissible stepsize parameters, some old and some new. For ease of reference, let us call \( \{D_{ii}^1\}_{i=1}^d \) the new parameters defined in (13), \( \{D_{ii}^2\}_{i=1}^d \) the existing one given in [3] (see (13)), \( \{D_{ii}^3\}_{i=1}^d \) the upper bound of \( \{D_{ii}^2\}_{i=1}^d \) used in [3] (see (17)), and \( \{D_{ii}^4\}_{i=1}^d \) the new upper bound of \( \{D_{ii}^2\}_{i=1}^d \) defined in (20).

In the next result we compare the four stepsizes. The lemma implies that \( D_4^1 \) and \( D_4^2 \) are uniformly smaller (i.e., better – see Theorem 3.2) than \( D_4^1 \) and \( D_4^3 \). However, \( D_4^2 \) involves quantities which are hard to compute (e.g., \( \sigma \)). Moreover, \( D_4^1 \) is always smaller than \( D_4^3 \).

**Lemma 6.1.** Let \( \tau \geq 2 \). The following holds for all \( i \):

\[ D_{ii}^1 \leq D_{ii}^2 \leq D_{ii}^3 \leq D_{ii}^4, \]  

(21)

\[ D_{ii}^1 \leq D_{ii}^4 \leq D_{ii}^3 \leq D_{ii}^2. \]  

(22)

**Proof.** We only need to show that \( D_{ii}^1 \leq D_{ii}^4 \), the other relations are already proved. It follows from (19) that

\[ \sum_{j=1}^n (\omega_j - \tilde{\sigma}) A_{ji}^2 \leq 0. \]  

(23)

holds for all \( i \). Moreover, letting

\[ \tilde{\sigma}' = \max_j \omega_j', \]  

(24)
and using Lemma 5.2 with $\eta = \bar{\sigma}$, we get
\[
(z - \frac{r-1}{s-1} \bar{\sigma}^{-1} \bar{\sigma} \leq \frac{1}{s-1} (1 + \frac{\bar{\sigma}(s-1)(r-1)}{s-1}).
\]
Now, for all $i$, we can write $D_{ii}^1 - D_{ii}^4 =
\begin{align*}
&\frac{\omega}{x} + \frac{\omega}{y} + \sum_{j=1}^{n} A_{ji}^2 \left( \frac{r-1}{s-1} (\omega_j - \bar{\sigma}) + \frac{r-1}{s-1} \omega_j^2 \right) \\
&\leq \frac{r-1}{s-1} \frac{1}{s-1} \left( (z - \frac{r-1}{s-1}) \left( \bar{\sigma}^{-1} \omega_j - \frac{1}{s-1} (1 + \frac{\bar{\sigma}(s-1)(r-1)}{s-1}) \right) \right) \\
&\leq \frac{r-1}{s-1} \frac{1}{s-1} \left( (z - \frac{r-1}{s-1}) \bar{\sigma}^{-1} \omega_j \right) \leq 0.
\end{align*}
\]

Note that we do not have a simple relation between $\{D_{ii}^1\}_{i=1}^d$ and $\{D_{ii}^2\}_{i=1}^d$. Indeed, for some coordinates $i$ the new stepsize parameters $\{D_{ii}^d\}_{i=1}^d$ could be smaller than $\{D_{ii}^d\}_{i=1}^d$, see Figure 1 in the next section for an illustration.

7. NUMERICAL EXPERIMENTS

In this section we present preliminary computational results involving real and synthetic data sets. In particular, we first compare the relative benefits of the 4 stepsize rules, and then demonstrate that Hydra can significantly outperform other methods on a big data non-strongly convex problem involving 50 billion variables.

The computational experiments were performed on two problems:

1. **Dual of SVM:** This problem can be formulated as finding $x \in [0,1]^d$ that minimizes
   \[
   L(x) = \frac{1}{2\lambda d^2} \sum_{j=1}^{n} \sum_{i=1}^{d} b_i A_{ji}^2 x_i^2 - \frac{1}{d} \sum_{i=1}^{d} x_i + I(x).
   \]
   Here $d$ is the number of examples, each row of the matrix $A$ corresponds to a feature, $b_i \in \mathbb{R}^d$, $\lambda > 0$ and $I(\cdot)$ denotes the indicator function of the set $[0,1]^d$:
   \[
   I(x) = \begin{cases} 
   0 & \text{if } x \in [0,1]^d, \\
   +\infty & \text{if } x \notin [0,1]^d.
   \end{cases}
   \]

2. **LASSO:**
   \[
   L(x) = \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1.
   \]

The computations were performed on Archer (http://archer.ac.uk/), which is UK’s leading supercomputer.

7.1. Benefit from the new stepsize

In this section we compare the four different stepsize parameters and show how they influence the convergence of Hydra. In the experiments of this section we used the astro-ph dataset\footnote{Astro-ph is a binary classification problem which consists of abstracts of papers from physics. The dataset consists of $d = 29,882$ samples and the feature space has dimension $n = 99,757$. Note that for the SVM dual problem, coordinates correspond to samples whereas for the LASSO problem, coordinates correspond to features.}. The problem that we solve is the SVM dual problem (for more details, see (10) [11]).

Figure 1 plots the values of $D_{ii}^1, D_{ii}^2, D_{ii}^3, D_{ii}^4$ for $i$ in $\{1,\ldots,d\}$ and $(c, \tau) = (32, 10)$. The dataset we consider is normalized (the diagonal elements of $A^T A$ are all equal to 1), and the parameters $\{D_{ii}^d\}_{i=1}^d$ are equal to a constant for all $i$ (the same holds for $\{D_{ii}^d\}_{i=1}^d$). Recall that according to Theorem 3.2 a faster convergence rate is guaranteed if smaller (but theoretically admissible) stepsize parameters $\{D_{ii}^d\}_{i=1}^d$ are used. Hence it is clear that $\{D_{ii}^d\}_{i=1}^d$ and $\{D_{ii}^d\}_{i=1}^d$ are better choices than $\{D_{ii}^d\}_{i=1}^d$ and $\{D_{ii}^d\}_{i=1}^d$. However, as we mentioned, computing the parameters $\{D_{ii}^d\}_{i=1}^d$ would require a considerable computational effort, while computing $\{D_{ii}^d\}_{i=1}^d$, is as easy as passing once through the dataset. For this relatively small dataset, the time is about 1 minute for computing $\{D_{ii}^d\}_{i=1}^d$ and less than 1 second for $\{D_{ii}^d\}_{i=1}^d$.

Fig. 1: Plots of $i$ vs $D_{ii}^1, i$ vs $D_{ii}^2, i$ vs $D_{ii}^3$ and $i$ vs $D_{ii}^4$

In order to investigate the benefit of the new stepsize parameters, we solved the SVM dual problem on the astro-ph dataset for $(c, \tau) = (32, 10)$. Figure 2 shows evolution of the duality gap, obtained by using the four stepsize parameters mentioned previously. We see clearly that smaller stepsize parameters lead to faster convergence, as predicted by Theorem 3.2. Moreover, using our easily computable new stepsize parameters $\{D_{ii}^d\}_{i=1}^d$, we achieve comparable convergence speed with respect to the existing parameters $\{D_{ii}^d\}_{i=1}^d$.

7.2. Hydra vs Hydra

In this section, we report experimental results comparing Hydra with Hydra on a synthetic big data LASSO problem.
We generated a sparse matrix $A$ having the same block angular structure as the one used in [3] Sec 7., with $d = 50$ billion, $c = 256$, $s = 195,312,500$ (note that $d = cs$) and $n = 5,000,000$. The average number of nonzero elements per row of $A$ (i.e., $\sum j \omega_j/n$) is 60,000, and the maximal number of nonzero elements in a row of $A$ (i.e., $\max j \omega_j$) is 1,993,419. The dataset size is 5TB.

We have used 128 physical nodes of the ARCHER supercomputing facility, which is based on Cray XC30 compute nodes connected via Aries interconnect. On each physical node we have run two MPI processes – one process per NUMA (NonUniform Memory Access) region. Each NUMA region has 2.7 GHz, 12-core E5-2697 v2 (Ivy Bridge) series processor and each core supports 2 hardware threads (Hyperthreads). In order to minimize communication we have chosen $\tau = s/1000$ (hence each thread computed an update for 8,138 coordinates during one iteration, on average).

We show in Figure 3 the decrease of the error with respect to the number of iterations, plotted in log scale. It is clear that Hydra provides faster iteration convergence rate than Hydra. Moreover, we see from Figure 4 that the speedup in terms of the number of iterations is significantly large so that Hydra converges faster in time than Hydra even though the run time of Hydra per iteration is on average two times more expensive.

8. REFERENCES


