Journal of Machine Learning Research 17 (2016) 1-24 Submitted 10/15; Revised 8/16; Published 12/16

Linear Convergence of Randomized Feasible Descent Methods Under the Weak Strong Convexity Assumption

Chenxin Ma CHM514@LEHIGH.EDU
Industrial and Systems Engineering
Lehigh University
Bethlehem, PA 18015, USA

Rachael Tappenden RACHAEL.TAPPENDEN@CANTERBURY.AC.NZ
Mathematics and Statistics
University of Canterbury
Christchurch 8041, New Zealand

Martin Takáč TAKAC.MT@GMAIL.COM
Industrial and Systems Engineering
Lehigh University
Bethlehem, PA 18015, USA

Editor: Leon Bottou

Abstract

In this paper we generalize the framework of the Feasible Descent Method (FDM) to a Randomized (R-FDM) and a Randomized Coordinate-wise Feasible Descent Method (RC-FDM) framework. We show that many machine learning algorithms, including the famous SDCA algorithm for optimizing the SVM dual problem, or the stochastic coordinate descent method for the LASSO problem, fits into the framework of RC-FDM. We prove linear convergence for both R-FDM and RC-FDM under the weak strong convexity assumption. Moreover, we show that the duality gap converges linearly for RC-FDM, which implies that the duality gap also converges linearly for SDCA applied to the SVM dual problem.

Keywords: feasible descent method, stochastic methods, iteration complexity, convergence theory, weak strong convexity

1. Introduction

In this paper we are interested in the following optimization problem

$$\min_{x \in X} f(x), \tag{1}$$

where the function f is smooth and convex, and $X \subseteq \mathbb{R}^n$ is a convex set. The Feasible Descent Method (FDM) (Luo and Tseng 1993; Necoara 2015; Wang and Lin 2014) is any algorithm, which produces a sequence of points $\{x_k\}_{k=0}^{\infty}$, where there exist constants $\beta \geq 0$, $\zeta > 0$ and $\omega_k \geq \bar{\omega} > 0$, such that for every iteration k, the following conditions are satisfied:

$$x_{k+1} = \mathbf{Proj}_X \left(x_k - \omega_k \nabla f(x_k) + z_k \right), \tag{2}$$

$$\|z_k\| \leq \beta \|x_k - x_{k+1}\|, \tag{3}$$

$$f(x_{k+1}) \leq f(x_k) - \zeta \|x_k - x_{k+1}\|^2, \tag{4}$$

©2016 Chenxin Ma, Rachael Tappenden and Martin Takáč.

where $\mathbf{Proj}_X(y) := \arg\min_{x\in X} \|x-y\|$ is the projection of y onto X.

As was shown in Luo and Tseng (1993), many first order algorithms, including steepest descent, the gradient projection algorithm, the extra gradient method, the proximal minimization algorithm and the cyclic coordinate descent method, fit into the framework of FDM. However, randomized first order algorithms are becoming increasingly popular in the optimization and machine learning literature, and the following question naturally arises:

"Can the framework of FDM be extended to a randomized setting?"

In this paper we give an affirmative answer to this question: we show that, indeed, a randomized version of FDM can be formulated and we will show that, for example, the inexact gradient projection algorithm (when the gradient is corrupted with random noise) or the stochastic coordinate descent method, fit into this new framework.

1.1 Assumptions and Notations

In this section we state the assumptions and introduce the notation that will be used in this paper. In particular, throughout this paper we will assume that f satisfies weak strong convexity, and that the gradient of f is Lipschitz.

Now we formalize the first assumption, which is that the function f enjoys weak strong convexity. Henceforth, we use $\mathbb{R}^n_{++}$ to denote the set of vectors in $\mathbb{R}^n$, with (strictly) positive components, and we denote the i-th component of the vector x by $x^{(i)}$.

Assumption 1. *We assume that there exists a positive vector $w \in \mathbb{R}^n_{++}$ such that the function $f(x)$ satisfies the* weak strong convexity property on the set X*, which is defined as*

$$f(x) - f(\bar{x}) \geq \kappa_f \|x-\bar{x}\|_W^2, \quad \forall x \in X, \tag{5}$$

where $\kappa_f > 0$, $W = \mathbf{diag}(w)$, $\|x\|_W^2 = \sum_{i=1}^n w_i (x^{(i)})^2$, $f^* = \min_{x\in X} f(x)$, and

$$\bar{x} := \arg \min_{y\in X: f(y)=f^*} \|x-y\|_W. \tag{6}$$

Let us remark that, if f is smooth and has a Lipschitz continuous gradient, then Assumption 1 is weaker than the *strong convexity* assumption or the *global error bound property*, Necoara (2015). We now provide a few examples of a functions that are used in machine learning, which are weakly strongly convex, but are not strongly convex.

1. In particular, let $x \in \mathbb{R}^n$ and let $c \in \mathbb{R}$. We have

   $$\text{shrink}_c(x) = \text{sign}(x)\max\{|x|-c, 0\}, \quad \text{and} \quad f(x) = \tfrac{1}{2}\|\text{shrink}_c(x)\|_2^2,$$

   where the shrinkage function is applied component-wise to vector x. Note that f is not strongly convex because $f(x) = 0$ for $x \in [-c, c]$ (which is the minimizer set). On the other hand, $f(x) = \frac{1}{2}\|x+c\|_2^2$ for $x \leq -c$ and $f(x) = \frac{1}{2}\|x-c\|_2^2$ for $x \geq c$. Thus, $f(x)$ is weakly strongly convex. See also Zhang and Yin (2013).

2. Another ilustrative example of function which is weakly stronly convex but not strongly convex is $f(x) = \frac{1}{2}\|Ax-b\|^2$ with $A \in \mathbb{R}^{m\times n}$ such that $m < n$. If x^* is some optimal solution then $x^* + t$ is also optimal iff $t \in \text{null}(A)$. One can easily show that κ_f is related to the smallest non-negative singular value of matrix $A^T A$.

The second assumption we make regards the smoothness of f, and is defined precisely as follows.

Assumption 2. *We assume that $f(x)$ has a coordinate-wise Lipschitz continuous gradient with constants L_i, i.e. $\forall x \in X$ and $\forall \delta \in \mathbb{R} : x + \delta e_i \in X$ the following inequality holds*

$$|\nabla_i f(x) - \nabla_i f(x + \delta e_i)| \leq L_i |\delta|, \tag{7}$$

where e_i denotes the i-th column of the identity matrix $I \in \mathbb{R}^{n \times n}$.

As was shown in Richtárik and Takáč (2014), Assumption 2 implies that the function $f(x)$ has a Lipschitz continuous gradient with Lipschitz constant $L_f^W > 0$ with respect to the norm $\|\cdot\|_W$, i.e. $\forall x, y \in X$ we have

$$\|\nabla f(x) - \nabla f(y)\|_W^* \leq L_f^W \|x - y\|_W, \tag{8}$$

where $\|x\|_W^* = \sqrt{\sum_{i=1}^n \frac{1}{w_i}(x^{(i)})^2}$ is the dual norm to $\|\cdot\|_W$. Moreover, Richtárik and Takáč (2014) also showed that $L_f^W \leq \sum_{i=1}^n \frac{L_i}{w_i}$.

We define the projection operator onto the set X, with respect to the norm $\|\cdot\|_W$, as follows

$$\mathbf{Proj}_X^W(x) = \arg\min_{y \in X} \|x - y\|_W^2 = \arg\min_{y \in X} \sum_{i=1}^n w_i (x^{(i)} - y^{(i)})^2. \tag{9}$$

1.2 Applications

In this section we discuss several problems that arise in the optimization and machine learning literature, which fit into the FDM framework that we analyze in this paper. We also provide details showing that, for each problem, the objective function satisfies the assumptions in Section 1.1. (A discussion regarding the value of the weak strong convexity parameter κ_f will be given in Section 4.)

The dual of SVM. Consider the classical linear SVM problem. The goal is, given n training points (a_i, y_i), where $a_i \in \mathbb{R}^d$ are the features for point i and $y_i \in \{-1, +1\}$ is its label, find $w \in \mathbb{R}^d$ such that the regularized empirical loss function is minimized, i.e., solve the following optimization problem

$$\min_{w \in \mathbb{R}^d} \left\{ \mathbf{P}(w) := \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell_i(w^T a_i) + \frac{\lambda}{2} \|w\|^2 \right\}, \tag{10}$$

where $\lambda > 0$ is a regularization parameter, and, in the case of SVM, the function $\ell_i(w^T a_i) = \max\{0, 1 - y_i w^T a_i\}$ is the hinge loss. Clearly, the objective function (10) is not smooth. However, one can formulate the dual problem (Hsieh et al. 2008; Shalev-Shwartz and Zhang 2013; Takáč et al. 2013)

$$\min_{x \in \mathbb{R}^n, 0 \leq x^{(i)} \leq 1} \left\{ f(x) := \tfrac{1}{2\lambda n^2} x^T Q x - \tfrac{1}{n} \mathbf{1}^T x \right\}, \tag{11}$$

where $Q_{i,j} = y_i y_j \langle a_i, a_j \rangle$, and $\mathbf{1}$ denotes the vector of all ones, which *is* smooth. The linear SVM problem (10) can now be solved via the dual problem (11). Note that (11) is

of the form (1), so our new FDM framework can be used to solve this important machine learning problem.

Lasso problem and least squares problem. Consider the following optimization problem

$$\min_{x\in\mathbb{R}^n} g(x) + \lambda\|x\|_1, \tag{12}$$

where $\lambda \geq 0$ and $g(x)$ is a smooth function with the special structure: $g(x) = h(Ax) + q^T x$, where $A \in \mathbb{R}^{m\times n}$ is some data matrix, $q \in \mathbb{R}^n$ is some vector and h is a strongly convex function. It is a simple exercise to show that, if we double the dimension of x to $[x^+; x^-]$, we can replace the term $\lambda\|x\|_1$ in (12) with $\lambda \mathbf{1}^T x^+ + \lambda \mathbf{1}^T x^-$ and impose the constraints $x^+, x^- \geq \mathbf{0}$. Then the Lasso problem (12) can be reformulated as a smooth optimization problem with simple box constraints.

ℓ_2 regularized empirical loss minimization. Many machine learning problems have the following structure (Chang et al. (2008))

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(a_i^T x) + \frac{\lambda}{2} x^T x, \tag{13}$$

where $\lambda > 0$ is a regularization parameter and ℓ_i is a loss function. Because we assume that f must be smooth, the following commonly used loss functions fit our assumptions: the logistic loss function $\ell_i(a_i^T x) = \log(1+\exp(-y_i a_i^T x))$; the squared loss function $\ell_i(a_i^T x) = (y_i - a_i^T x)^2$ and the squared hinge loss function $\ell_i(a_i^T x) = (\max\{0, 1 - y_i a_i^T x\})^2$. Hence, any machine learning problem of the form (13) (used with any of the mentioned loss functions) fits our randomized FDM framework.

1.3 Related work

Luo and Tseng (1993) are among the first to establish asymptotic linear convergence for a non-strongly convex problem under the local error bound property. They consider a class of feasible descent methods, which includes, for example, the cyclic coordinate descent method. The error bound measures how close the current solution is to the optimal solution set, with respect to the projected gradient. Recently, Wang and Lin (2014) proved that the feasible descent method enjoys a linear convergence rate (from the beginning, rather than only locally) under the global error bound property. Considering the class of smooth constrained optimization problems with the global error bound property, Necoara and Clipici (2016); Necoara and Nedelcu (2014a) showed a linear convergence rate for the parallel version of the stochastic coordinate descent method. Liu and Wright (2015) analyzed the asynchronous stochastic coordinate descent method (SCDM) under the weak strong convexity assumption. Very recently, Necoara (2015) showed that, if the objective function is smooth, then the class of problems with the global error bound property is a subset of the class of problems with the weak strong convexity property.

1.4 Contributions

In this section we list the most important contributions of this paper (not in order of their significance):

- **Randomized and Randomized Coordinate Feasible Descent Methods.** We extend the well known framework of Feasible Descent Methods (FDM) (Luo and Tseng 1993) to randomized and randomized coordinate FDM and show that the SCDM algorithm fits into our new proposed framework.

- **Linear Convergence Rate.** We show that any stochastic or deterministic algorithm, which fits our Randomized FDM (R-FDM) or Randomized Coordinate-FDM (RC-FDM) framework and satisfies our previously stated assumptions, converges linearly in expectation.

- **Linear Convergence of the Duality Gap for SDCA for SVM**. As a consequence of our analysis, we show that when SDCA is applied to the dual of the SVM problem, the duality gap converges linearly. Previously, linear convergence of the duality gap was only proven in case when the matrix Q in (11) is positive definite (Shalev-Shwartz and Zhang 2013; Takáč et al. 2015). However, our new linear convergence result holds, even when Q is singular.

- **Inexact Randomized Coordinate Descent.** By the nature of the FDM framework, *inexact* first order methods belong to the class of FDMs, (where inexact methods are methods that incorporate some kind of inexact information, for example, via inexact gradients, or via inexact updates). Our new randomized coordinate FDM framework includes inexact randomized coordinate descent methods. Therefore, another contribution of this work is that it provides a linear convergence rate for e.g. randomized coordinate descent with *inexact* computations of (partial) gradient, which was analyzed in various settings. (See, for example, Devolder et al. 2014; Bonettini 2011; Tappenden et al. 2016; Hua and Yamashita 2012; Necoara and Nedelcu 2014b.)

- **Flexibility and wide applicability.** Our randomized- and randomized coordinate-FDM framework is extremely *flexible*. It is a general framework that not only covers and unifies many existing algorithms, but any algorithm that fits our framework is also covered by the FDM convergence guarantees. Moreover, as demonstrated in Section 1.2, a very wide range of optimization and machine learning problems can be written in the form (1), and subsequently, they can be solved via the new FDM framework. Problems include the dual of SVM, the LASSO problem, and any ℓ_2-regularized empirical loss functions where the loss function is smooth and separable. All such problems appear very frequently in the machine learning literature.

- **Parallel methods.** The RC-FDM framework is sufficiently general so as to include parallel randomized coordinate descent methods.

1.5 Paper Outline

In Section 2 we derive the Randomized (R-FDM) and the Randomized Coordinate (RC-FDM) Feasible Descent Methods. In Section 3 we derive the convergence rate for any method which fits into the R-FDM or RC-FDM framework and we compare our results with those in Liu and Wright (2015) for SCDM. In Section 4 we briefly review the global error bound property and using the result in Necoara (2015) we compare our convergence

results with Wang and Lin (2014). In Section 5 we show that the duality gap converges linearly for SDCA applied to the dual of the SVM problem, and in Section 6 we present a brief summary.

2. Randomized and Randomized Coordinate Feasible Descent Method

The framework of Feasible Descent Methods (FDM) broadly covers many algorithms that use first-order information (Luo and Tseng 1993) including gradient descent, cyclic coordinate descent and also the inexact gradient descent algorithm. We generalize the classical FDM framework to a randomized setting, which we call the Randomized Feasible Descent Method (R-FDM). Algorithms that use randomization have become extremely popular over the past few years, and the success, reliability, scalability, applicability and efficiency of such random algorithms is well documented. To the best of our knowledge this is the first time such a unifying R-FDM framework has been proposed and that a global linear convergence rate has been established under Assumptions 1 and 2. Further, we also show that the popular minibatch stochastic coordinate descent/ascent method, fits into the R-FDM framework.

Definition 3 (Randomized Feasible Descent Method (R-FDM))**.** *A sequence* $\{x_k\}_{k=0}^{\infty}$ *is generated by R-FDM if there exist* $\beta \geq 0$, $\zeta > 0$ *and* $\{\omega_k\}_{k=0}^{\infty}$ *with* $\min_k \omega_k \geq \bar{\omega} > 0$ *such that for every iteration* k*, the following conditions are satisfied,*

$$x_{k+1} = \mathbf{Proj}_X^W \left(x_k - \omega_k W^{-1}(\nabla f(x_k) - z_k) \right), \tag{14}$$

$$\mathbf{E}[(\|z_k\|_W^*)^2] \leq \beta^2 \mathbf{E}[\|x_k - x_{k+1}\|_W^2], \tag{15}$$

$$\mathbf{E}[f(x_{k+1})] \leq f(x_k) - \zeta \mathbf{E}[\|x_k - x_{k+1}\|_W^2], \tag{16}$$

where z_k *is some random vector, which satisfies the Markov property, conditioned on* x_k*.*

We will now compare the new Randomized FDM framework (Definition 3) with the original FDM ((2)–(4)), where, for simplicity of exposition, we will take $\|\cdot\|_W \equiv \|\cdot\|_2$ (i.e., $W = I$). Notice that the first step of R-FDM (14) is the same as the first step of FDM (2); it is a projected (gradient) descent type step. Note the role that z_k plays in (14); it captures any error/inexactness/noise in the update step, and it is clear to see that if $z_k = 0$ for all k, (i.e., no inexactness) then (14) is the same update as in a projected gradient descent method. Next, (15) gives an upper bound for $\|z_k\|_W^*$, which shows that, in order to guarantee convergence, the noise in (14) cannot be arbitrarily large, which intuitively makes sense. Finally, (16) guarantees that there is a reduction in the objective value, in expectation, after each iteration. The key difference between FDM and R-FDM is that for FDM, (3) and (4) hold deterministically (with a deterministic vector z_k), whereas for R-FDM (3) and (4) only need to hold *in expectation*. That is, for R-FDM, conditions (3) and (4) are *replaced by conditions* (15) *and* (16), where z_k is a *random vector*. Notice that (15) and (16) are weaker conditions than (3) and (4). That is, for FDM, (3) and (4) must hold at every iteration (i.e., they are deterministic), whereas for the R-FDM framework, the conditions (15) and (16) are equivalent to (3) and (4) holding *only on average*. The R-FDM framework is *extremely general*. It encapsulates algorithms that involve a (possibly noisy) projection step, has small enough noise *on average* and decreases the objective function *on average*, as the iterations progress.

Remark 4. *We will see later (in the proof of convergence of R-FDM) that* (15) *can be relaxed to the existence of constant* $\eta > 0$ *such that* $\mathbf{E}[(\|z_k\|_W^*)^2] \leq \eta \left(f(x_k) - \mathbf{E}[f(x_{k+1})]_W^2\right)$.

We will now demonstrate that (see Theorem 7), under an additional mild assumption, if the set $X = \mathbb{R}^n$, then SCDM (captured in Algorithm 1 with Option I.) is equivalent to R-FDM. We also remark that there is a need to modify R-FDM so that the stochastic coordinate descent method can be analyzed even when $X \neq \mathbb{R}^n$. However, first we describe SCDM and make the following assumption in order to establish the equivalence of SCDM with $X = \mathbb{R}^n$ and R-FDM.

Algorithm 1 Stochastic Coordinate Descent Method (SCDM)

Input: $f(x)$, $\{\omega_k\}_{k=0}^{\infty}$, diagonal matrix $W \succ 0$, x_0.
Input: $X = X_1 \times \cdots \times X_n$, where $X_i = [a, b]$ with $-\infty \leq a < b \leq +\infty$
while $k \geq 0$: **do**
  choose $i \in \{1, 2, \ldots, n\}$ uniformly at random
  set $x_{k+1} = x_k$
  Option I: $x_{k+1}^{(i)} = \arg\min_{x^{(i)} \in X_i} f((x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, x^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T)$
  Option II: $x_{k+1} = \mathbf{Proj}_X^W \left(x_k - \omega_k W^{-1} \nabla_i f(x_k) e_i\right)$
end while

Remark 5. *For simplicity of exposition, Algorithm 1 is the serial form of SCDM, although a minibatch version of SCDM does exist. We will see in Definition 9 that our RC-FDM framework is flexible and general, because it also works in the parallel/minibatch case.*

Assumption 6. *The function* f *is coordinate-wise strongly convex with respect to the norm* $\|\cdot\|_W$ *with parameter* $\gamma > 0$*, if, for any* $x \in X$ *and any* $i \in \{1, 2, \ldots, n\}$ *we have*

$$f(x^{(1)}, \ldots, x^{(i-1)}, \xi, x^{(i+1)}, \ldots, x^{(n)}) - f(x) + \nabla_i f(x)(x^{(i)} - \xi) \geq \gamma w_i |\xi - x^{(i)}|^2. \qquad (17)$$

Note that Assumption 6 *does not imply strong convexity of the function* f. For example, (17) is satisfied for the Lasso problem or for the SVM dual problem whenever $\forall i : \|a_i\| > 0$, and neither of those problems is strongly convex.

Theorem 7. *Let Assumptions 1, 2 and 6 hold. If* $X = \mathbb{R}^n$ *then the Stochastic Coordinate Descent Method (SCDM) (Algorithm 1 with Option I.) is equivalent to R-FDM with the parameters* $\beta^2 = 2[(L_f^W)^2 + 1] + (n-1)r^2$, $\zeta = \gamma$ *and* $\omega_k = 1$*, where* $r^2 = \max_i \frac{L_i^2}{w_i^2}$.

Let us comment that, if importance sampling is incorporated into SCDM, the convergence rate in Theorem 7 can be slightly improved, as the parameter r can be made to be $r^2 = \frac{1}{n}\sum_i \frac{L_i^2}{w_i^2}$, i.e., 'average' rather than 'max'. However, it is typical to consider the case $w_i = L_i$ so that $r^2 = 1$ in Theorem 7 regardless.

The following remark compares the result of the above theorem with the cyclic rule.

Remark 8. *It was shown in Luo and Tseng (1993) that for the cyclic coordinate descent method (which is not randomized and hence Equation 14-16 hold deterministically) we have*

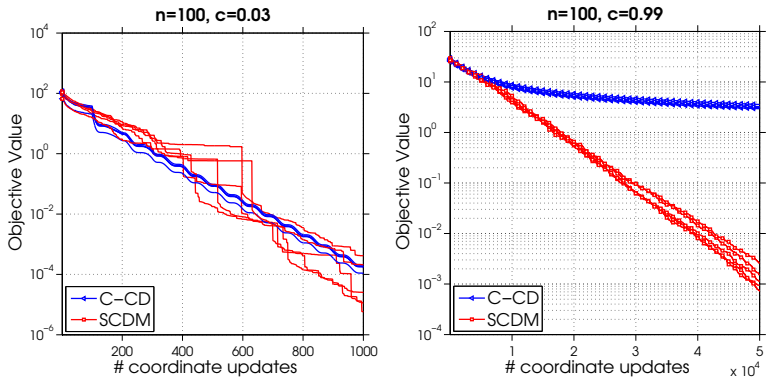

Figure 1: Number of coordinate updates v.s. objective gap for two methods.

$\omega_k^{cyclic} = 1$, $\zeta^{cyclic} = \gamma$ *and* $(\beta^{cyclic})^2 = (1+\sqrt{n}L_f^W)^2 = 1+2\sqrt{n}L_f^W + n(L_f^W)^2$. *For simplicity, let us assume that* $W = \mathbf{diag}(L_1, L_2, \ldots, L_n)$. *Then* $r^2 = 1$ *and* $L_f^W \in [1, n]$. *For the cyclic coordinate descent method and SCDM,* ω_k *and* ζ *are the same. However, if we consider the worst case (when* $L_f^W = n$*) we have that* $\beta^2 \sim \mathcal{O}(n^2)$*, whereas* $(\beta^{cyclic})^2 \sim \mathcal{O}(n^3)$. *Also note that one iteration of cyclic coordinate descent requires* n *coordinate updates, whereas SCDM updates just* **one** *coordinate, and therefore each iteration of SCDM is* n *times cheaper. In the other extreme, when* $L_f^W = 1$ *we have that both* $\beta^2 \sim (\beta^{cyclic})^2 \sim \mathcal{O}(n)$*, but again we recall that one iteration of SCDM is* n *times cheaper.*

We present two experiments to support the discussion above. We apply both the cyclic coordinate descent method and SCDM to the problem

$$\min_{x \in \mathbb{R}^n} f(x) = x^T A x,$$

where the matrix $A \in \mathbb{R}^{n \times n}$ has ones on the diagonal and constant c elsewhere, Sun and Ye (2016). The optimal solution to this problem is $x = 0$. For the first experiment we set $n = 100$ and $c = 0.03$ ($L_f^W = 3.97 \approx 1$), and in the second experiment we keep n unchanged and set $c = 0.99$ ($L_f^W = 99.01 \approx n$). For each method we randomly select five starting points. From Figure 1, it is easy to see that when c is large (i.e., $L_f^W \approx n$) SCDM performs much better than cyclic coordinate descent. On the other hand, there is not such an obvious difference between the two methods when c is small. Thus, the difference in performance between the two methods depends upon the parameter L_f^W. Moreover, the case where c is small is more friendly for both methods, since they require far fewer coordinate updates to reach optimality, compared with the large c case. These results highlight and support Remark 8, regarding the theoretical gap between two methods.

It turns out that if $X \neq \mathbb{R}^n$ then SCDM does not fit the R-FDM framework because $\nabla_i f(x_k)$ cannot be bounded by $\|x_k - x_{k+1}\|_W$, as is shown in the proof of Theorem 7. Thus, there is a need to modify R-FDM such that the SCDM algorithm can be analyzed for bounded problems.

The natural modification to R-FDM, which would allow SCDM to fit the R-FDM framework is the following: at each iteration k we require that in (14), only a subset of coordinates of the vector x_k are updated. This can be achieved by the following method.

Definition 9. *[Randomized Coordinate Feasible Descent Method (RC-FDM)] Let $X = X_1 \times \cdots \times X_n$, where X_i are intervals. A sequence $\{x_k\}_{k=0}^{\infty}$ is generated by RC-FDM if there exists $\beta \geq 0$, $\zeta > 0$ and $\{\omega_k\}_{k=0}^{\infty}$ with $\min_k \omega_k \geq \bar{\omega} > 0$ such that for every iteration k, the following are satisfied*

$$x_{k+1} = \mathbf{Proj}_X^W \left(x_k - \omega_k W^{-1} (\nabla f(x_k) - z_k)_{[\mathcal{I}]} \right), \tag{18}$$

$$(\|(z_k)_{[\mathcal{I}]}\|_W^*)^2 \leq \beta^2 \|x_k - x_{k+1}\|_W^2, \tag{19}$$

$$f(x_{k+1}) \leq f(x_k) - \zeta \|x_k - x_{k+1}\|_W^2, \tag{20}$$

where $\mathcal{I}$ is a set of coordinates that are selected uniformly at random from the set $\{1, 2, \ldots, n\}$ with $|\mathcal{I}| = \tau$, where $1 \leq \tau \leq n$, $x_{[\mathcal{I}]}$ is a vector whose elements $j \notin \mathcal{I}$ are set to 0 and z_k is some fixed vector at iteration k.

Now, we show that even if $X \neq \mathbb{R}^n$, SCDM fits the RC-FDM. Theorem 10 holds if Option I. is used in Algorithm 1 and Theorem 11 holds if Option II. is used.

Theorem 10. *Let Assumptions 1, 2 and 6 hold. Let $\tau = 1$ for simplicity, if $X = X_1 \times \cdots \times X_n$, where X_i are intervals then the Stochastic Coordinate Descent Method in Algorithm 1 with Option I. is RC-FDM with $\beta^2 = 2[(L_f^W)^2 + 1]$, $\zeta = \gamma$, and $\omega_k = 1$.*

Theorem 11. *Let Assumptions 1, 2 and 6 hold. Let $\tau = 1$ for simplicty, if $X = X_1 \times \cdots \times X_n$, where X_i are intervals then the Stochastic Coordinate Descent Method in Algorithm 1 with Option II. is RC-FDM with $z_k = 0$, $\zeta = \gamma$, $\beta = 0$, $\omega_k = 1$, and $W = \mathbf{diag}(L_1, L_2, \ldots, L_n)$.*

3. Convergence Analysis

Necoara (2015) proved a linear convergence rate for FDM under Assumptions 1 and 2. The following theorem shows that a linear convergence rate can also be established for R-FDM.

Theorem 12 (Linear Convergence of R-FDM)**.** *Let Assumptions 1 and 2 hold. If the sequence $\{x_k\}_{k=0}^{\infty}$ is produced by R-FDM, i.e.* (14)-(16) *are satisfied, then*

$$\mathbf{E}[f(x_k) - f^*] \leq \left(\frac{c}{1+c} \right)^k (f(x_0) - f^*), \tag{21}$$

where

$$c = \frac{2}{\kappa_f \zeta} \left((L_f^W + \tfrac{1}{\bar{\omega}})^2 + \beta^2 \right). \tag{22}$$

The next theorem establishes a linear convergence rate for RC-FDM.

Theorem 13 (Linear Convergence of RC-FDM)**.** *Let $X = X_1 \times \cdots \times X_n$, where X_i are intervals. Further, let Assumptions 1 and 2 hold, and let the sequence $\{x_k\}_{k=0}^{\infty}$ be produced by RC-FDM, i.e.* (18)-(20) *are satisfied. Then, for $z_k \neq 0$, there exists $c \in (0, 1)$ such that, for all k,*

$$\mathbf{E}[f(x_k) - f^*] \leq (1 - c)^k (f(x_0) - f^*). \tag{23}$$

Moreover, if for all k we have $z_k \equiv 0$, and $\frac{1}{\omega_k} \geq \max_i \frac{L_i}{w_i}$, then $c = \frac{2\bar{\omega}\tau}{n(2\bar{\omega}+1)}$ with

$$\mathbf{E}[f(x_k) - f^*] \leq (1 - c)^k \left(f(x_0) - f^* + \frac{\tau}{2\bar{\omega}} \|x_0 - \bar{x}_0\|_W^2 \right). \tag{24}$$

3.1 Comparison with the Results in Related Literature

In Theorem 13 we established a linear rate of convergence for RC-FDM for any z_k. We will now compare our result with the one presented in Liu and Wright (2015) for the projected coordinate gradient descent algorithm, and also with the result presented in Necoara (2015) for deterministic FDM. For this comparison we will assume that $\tau = 1$ (i.e., for serial RC-FDM), because the first paper only considers a serial algorithm, and for ease of comparison with the results in the second paper. (However, RC-FDM works for general $1 \leq \tau \leq n$.)

The projected coordinate gradient descent algorithm (Liu and Wright 2015) fits the RC-FDM framework. We also note that the result in Liu and Wright (2015) only holds for $z_k = 0$, so our result is more general. Further, even though the paper Liu and Wright (2015) considers an asynchronous implementation, where the update computed at iteration k is based on gradient information at a point up to ν iterations old, if $\nu = 0$ then their method fits into the RC-FDM framework. One of the benefits of our work is that more general norms can be used. So, for simplicity, and to match with the work in Liu and Wright (2015), let us assume that $L_i = 1$ for all i and we also choose $w_i = 1$ for all i. (This is the case e.g. for the SVM dual problem). The geometric rate in (24) in our work is then $1 - \frac{\kappa_f}{n(\kappa_f + \frac{1}{2})}$ and from Theorem 4.1 in Liu and Wright (2015) for $\nu = 0$ we obtain that the geometric rate is $1 - \frac{\kappa_f}{n(\kappa + L_{\max})}$, where $L_{\max} \geq 1$ is such that

$$\|\nabla f(x) - \nabla f(x + \delta e_i)\|_\infty \leq L_{\max}|\delta|$$

holds $\forall x \in \mathbb{R}^n$, $\delta \in \mathbb{R}$ and $i \in \{1, 2, \ldots, n\}$. Hence, in this case our convergence results are better because $\frac{1}{2} \leq 1 \leq L_{\max}$.

In Necoara (2015) the author provided a linear convergence rate for deterministic FDM. It is shown in Theorem 3.2 in Necoara (2015) that the coefficient of the linear rate is $1 - \frac{\zeta}{\zeta + \rho}$ where $\rho = \frac{1}{\kappa_f}(L_f + \frac{1}{\bar{\omega}} + \beta)^2$ whereas, in Theorem 13 of this work, from (21) we see that the coefficient is the same but with a different ρ. To be precise, in our case we have $\bar{\rho} = \frac{2}{\kappa_f}\left((L_f^W + \frac{1}{\bar{\omega}})^2 + \beta^2\right)$. Our result can be better or worse than that in Necoara (2015), depending on the values of $L_f^W, \bar{\omega}$ and β, but our results holds for R-FDM, which is broader than FDM.

4. Global Error Bound Property

In this section we describe a class of problems that satisfies the Global Error Bound (GEB) property. We show that this implies the weak strong convexity property and we compare the convergence rate obtained in this paper with several results in the current literature derived for problems obeying the GEB. We begin by defining the projected gradient.

Definition 14 (Projected Gradient)**.** *For any $x \in \mathbb{R}^n$ let us define the projected gradient as follows,*

$$\nabla^+ f(x) := x - \mathbf{Proj}_X^W\left(x - \nabla f(x)\right). \tag{25}$$

Note that projected gradient is zero at x if and only if x is an optimal solution of (1). Also, we will employ the projected gradient to define an error bound, which measures the distance between x and the optimal solution. Now, we are ready to define a *global error bound* as follows.

Definition 15 (Definition 6 in Wang and Lin 2014)**.** *An optimization problem admits a global error bound if there is a constant $\eta_f \geq 0$ such that*

$$\|x - \bar{x}\| \leq \eta_f \|\nabla^+ f(x)\|_W^*, \quad \forall x \in X, \tag{26}$$

where $\bar{x}$ and $\nabla^+ f(x)$ are defined in (6) *and* (25)*, respectively. A relaxed condition called* the global error bound from the beginning *is if the above inequality holds only for $x \in X$ such that $f(x) - f(\bar{x}) \leq M$, where M is a constant, and usually we have that $M = f(x_0) - f^*$.*

Let us consider a special instance of (1) when X is a polyhedral set, i.e.

$$X = \{x \in \mathbb{R}^n : Bx \leq c\}, \tag{27}$$

and the function f has the following structure

$$f(x) = h(Ax) + q^T x, \tag{28}$$

where $B \in \mathbb{R}^{l \times n}$, $A \in \mathbb{R}^{d \times n}$, h is a σ_h strongly convex function and f satisfies Assumption 2. We also assume that there exists an optimal solution and hence the optimal solution set X^* is assumed to be non-empty, Wang and Lin (2014). It is easy to observe that if f is strongly convex, then (5) is trivially satisfied. Just recently, Necoara (2015) showed that if (26) is satisfied, then (5) is satisfied with

$$\kappa_f = \frac{L_f^W}{2\eta_f^2}. \tag{29}$$

For problem (28) it was discussed in Wang and Lin (2014) that

$$\eta_f = \theta^2(1 + L_f^W)\left(\frac{1 + 2\|\nabla h(A\bar{x})\|^2}{\sigma_h} + 4M\right) + 2\theta\|\nabla f(\bar{x})\|, \tag{30}$$

where θ is a constant from the Hoffman bound (Hoffman 1952; Li 1993; Robinson 1973) defined as follows

$$\theta := \sup_{u,v} \left\{ \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\| \;\middle|\; \begin{array}{l} \left\| B^T u + \begin{pmatrix} A \\ q^T \end{pmatrix}^T v \right\| = 1, u \geq 0 \\ \text{and the corresponding rows of } B, A \text{ to } u, v\text{'s} \\ \text{non-zero elements are linearly independent.} \end{array} \right\}. \tag{31}$$

Note that the constant θ can be very large; we will discuss this in Section 5.

Necoara (2015) derived that, for problem (28), the weak strong convexity property (5) holds with

$$\kappa_f = \frac{\sigma_h}{2\theta^2}. \tag{32}$$

Note that κ_f given in (32) is $\mathcal{O}(\theta^{-2})$ whereas κ_f obtained from (29) is of the order θ^{-4}. Therefore we will compare our results using the latter estimates of κ_f.

4.1 Comparison with the Results in Related Literature

In Theorem 8 in Wang and Lin (2014), under the global error bound property, it is proven that FDM converges at a linear rate: $f(x_{k+1}) - f^* \leq (1 - \frac{1}{\bar{c}+1})(f(x_k) - f^*)$, with[1]

$$\bar{c} = \frac{1}{\zeta}(L_f^W + \frac{1}{\bar{\omega}} + \beta)(1 + \eta_f(\frac{1}{\bar{\omega}} + \beta)) = \frac{1}{\zeta}(L_f^W + \frac{1}{\bar{\omega}} + \beta)(1 + \theta^2 \frac{1 + L_f^W}{\sigma_h}(\frac{1}{\bar{\omega}} + \beta))$$
$$\sim \mathcal{O}\left(\frac{\theta^2}{\zeta \sigma_h}(1 + L_f^W)(\frac{1}{\bar{\omega}} + \beta)(L_f^W + \frac{1}{\bar{\omega}} + \beta)\right).$$

From Theorem 12 in this work, we have linear convergence of RC-FDM with the coefficient

$$c = \frac{2}{\kappa_f \zeta}\left((L_f^W + \tfrac{1}{\bar{\omega}})^2 + \beta^2\right) \overset{(32)}{=} \frac{4\theta^2}{\sigma_h \zeta}\left((L_f^W + \tfrac{1}{\bar{\omega}})^2 + \beta^2\right).$$

These coefficients are very similar, but FDM Wang and Lin (2014) covers only cyclic coordinate descent and not a randomized coordinate descent method (which is covered by Theorem 12).

5. Linear Convergence Rate of SDCA for Dual of SVM

In this section we show that the SDCA algorithm (which is SCDM applied to Equation 11) achieves a linear convergence rate for the duality gap. This improves upon the result obtained in Shalev-Shwartz and Zhang (2013); Takáč et al. (2015); Takáč et al. (2013), where only a sublinear rate was derived.

Assume, for simplicity, that in problem (10) for all $i \in \{1, 2, \ldots, n\}$ it holds that $\|a_i\| \leq 1$. Then, from Takáč et al. (2015); Takáč et al. (2013), we have that for any $x \in \mathbb{R}^n, s \in [0, 1]$ and the function f defined in (11),

$$f(x) - f^* \geq sG(x) - s^2 \frac{\sigma^2}{2\lambda}, \tag{33}$$

where f^* denotes the optimal value of (11), $A = [a_1, a_2, \ldots, a_n], \sigma^2 = \frac{1}{n}\|X\| \in [\frac{1}{n}, 1]$ and $G(x)$ is the duality gap at the point x, which is defined as $G(x) := P(\frac{1}{\lambda n} Ax) + f(x)$.

We remark that SDCA for problem (11) is equivalent to RC-FDM, where the constants in (18)-(20) are: $z_k = 0$, $\beta^2 = 0$, $w_i = L_i = \frac{1}{\lambda n^2}\|a_i\|^2$, and $\omega_k = 1$. Hence, if we choose $x_0 = \mathbf{0}$ then from Theorem 13 we have that $\mathbf{E}[f(x_k) - f^*] \leq (1 - c)^k \left(f(\mathbf{0}) - f^* + \|x^*\|_L^2\right)$ with $c = \frac{2\kappa_f}{n(2\kappa_f + 1)}$.

Now, we see that rearranging (33) gives

$$G(x) \overset{(33)}{\leq} s\frac{\sigma^2}{2\lambda} + \frac{1}{s}(f(x) - f^*). \tag{34}$$

If we want to achieve $G(x) \leq \epsilon$ it is sufficient to choose both terms on right hand side of (34) to be $\leq \frac{\epsilon}{2}$. Hence, we can set $s = \min\{1, \frac{\epsilon\lambda}{\sigma^2}\}$. All we have to do now is to choose k such that $f(x_k) - f^* \leq s\frac{\epsilon}{2}$. In the following theorem we establish linear convergence of the duality gap $G(x)$ for the SDCA algorithm.

1. In Wang and Lin (2014) it is shown that, in special cases (e.g. $X = \mathbb{R}^n$), (30) is $\eta_f = \theta^2 \frac{1 + L_f^W}{\sigma_h}$.

Theorem 16. *Let* $s = \min\{1, \frac{1}{\epsilon\lambda}\sigma^2\}$ *and let* K *be such that*

$$K \geq n\left(1 + \frac{1}{2\kappa_f}\right)\log\frac{2\left(f(\mathbf{0}) - f^* + \|x^*\|_L^2\right)}{s\epsilon}.$$

Then if the SDCA algorithm is applied to problem (11) *to produce* $\{x_k\}_{k=0}^{\infty}$*, then* $\forall k \geq K$ *we have that* $\mathbf{E}[G(x_k)] \leq \epsilon$*.*

Let us now comment on the size of the parameter $\kappa_f \overset{(32)}{=} \frac{\sigma_h}{2\theta^2}$. In our case, X is the polyhedral set (27) defined by $B = \begin{pmatrix}-I_n, & I_n\end{pmatrix}^T$, and $c = (\mathbf{0}^T, \mathbf{1}^T)^T$, where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix. Because of this structure (31) simplifies to

$$\theta := \sup_{u,n}\left\{\left\|\begin{pmatrix}u\\v\end{pmatrix}\right\| \;\middle|\; \begin{array}{l}\left\|I_n u + \begin{pmatrix}A\\q^T\end{pmatrix}^T v\right\| = 1\\ \text{and the corresponding rows of } I_n, A \text{ to } u, v\text{'s}\\ \text{non-zero elements are linearly independent.}\end{array}\right\}. \tag{35}$$

To show that θ can be very large, let us assume that two rows of the matrix A are highly correlated (in this case rows corresponds to features). We denote these two rows by A_1 and A_2, and let us assume that $A_1 = A_2 + \delta e_1$. Then we can chose $v = (-\frac{1}{\delta}, \frac{1}{\delta}, 0, \ldots, 0)^T$ and $u = \mathbf{0}$. This particular choice is feasible in optimization problem (35) and hence is imposing a lower-bound on θ: $\theta \geq \frac{\sqrt{2}}{|\delta|}$. Clearly, for small δ, this shows that θ can be arbitrarily large.

6. Summary

In this paper we have extended the framework of the feasible descent method FDM to a randomized, and a randomized coordinate, FDM framework. We have shown that many problems in the machine learning literature fit our problem structure, and subsequently, any algorithm that fits our FDM framework can be used to successfully solve them. We have proven a linear convergence rate (under the weak strong convexity assumption) for both methods, and we have shown that the convergence rates are similar to the deterministic/non-randomized FDM. We also showed that for the cyclic coordinate descent method, the coefficients in FDM are worse than, or similar to, the stochastic coordinate descent method (and hence the theory tells us that they converge at roughly the same speed), but each iteration of the stochastic coordinate descent method is n-times cheaper. We concluded the paper with a result showing that, for the SDCA algorithm applied to the dual of the linear SVM, the duality gap converges linearly.

Acknowledgments

Chenxim Ma and Martin Takáč were supported by National Science Foundation grant CCF-1618717.

Appendix A. Proof of Theorem 7

Let us define an auxiliary vector $\tilde{x}$ such that

$$\tilde{x}^{(i)} = \arg\min_{x^{(i)} \in X_i} f((x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, x^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T). \tag{36}$$

Then we can see that if coordinate i is chosen during iteration k in Algorithm 1 then

$$x_{k+1}^{(j)} = \begin{cases} x_k^{(j)}, & \text{if } \ j \neq i, \\ \tilde{x}^{(i)}, & \text{otherwise.} \end{cases} \tag{37}$$

If coordinate i is chosen during iteration k, then the optimality conditions for Step 6 of Algorithm 1, give us that

$$x_{k+1}^{(i)} = \mathbf{Proj}_{X_i}^W \left(x_{k+1}^{(i)} - \frac{1}{w_i} \nabla_i f(x_{k+1}) \right). \tag{38}$$

Moreover, by (37), for $j \neq i$ we have that $x_k^{(j)} = x_{k+1}^{(j)}$ which is possible only if $z_k^{(j)} = \nabla_j f(x_k)$.

Note that x_{k+1} is a random variable, which depends on i and x_k only. Therefore, we can define a random z_k such that the i-th coordinate is

$$z_k^{(i)} = \nabla_i f(x_k) - \nabla_i f((x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T) + w_i(x_k^{(i)} - \tilde{x}^{(i)}) \tag{39}$$

and the j-th coordinate (for $j \neq i$) is defined as $z_k^{(j)} = \nabla_j f(x_k)$. It is easy to verify that for z_k defined above, condition (14) holds. Now, we will compute $\mathbf{E}[(\|z_k\|_W^*)^2]$. We have that if the i-th coordinate is chosen then

$$\begin{aligned}
&\frac{1}{w_i}(z_k^{(i)})^2 \\
&= \frac{1}{w_i}\Big(\nabla_i f(x_k) - \nabla_i f((x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T) + w_i(x_k^{(i)} - \tilde{x}^{(i)})\Big)^2 \\
&\leq \frac{2}{w_i}\Big(\nabla_i f(x_k) - \nabla_i f((x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T)\Big)^2 + 2w_i(x_k^{(i)} - \tilde{x}^{(i)})^2 \\
&\leq 2\big(\|\nabla f(x_k) - \nabla f((x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T)\|_W^*\big)^2 + 2w_i(x_k^{(i)} - \tilde{x}^{(i)})^2 \\
&\overset{(8)}{\leq} 2(L_f^W \|x_k - (x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T\|_W)^2 + 2w_i(x_k^{(i)} - \tilde{x}^{(i)})^2 \\
&= 2(L_f^W)^2 w_i(x_k^{(i)} - \tilde{x}^{(i)})^2 + 2w_i(x_k^{(i)} - \tilde{x}^{(i)})^2 = 2[(L_f^W)^2 + 1]w_i(x_k^{(i)} - \tilde{x}^{(i)})^2,
\end{aligned} \tag{40}$$

otherwise

$$\frac{1}{w_i}(z_k^{(i)})^2 = \frac{1}{w_i}(\nabla_i f(x_k))^2.$$

Hence, we obtain that

$$\mathbf{E}[(\|z_k\|_W^*)^2] \overset{(40)}{\leq} \sum_{i=1}^n \frac{1}{n} 2[(L_f^W)^2 + 1]w_i(x_k^{(i)} - \tilde{x}^{(i)})^2 + \frac{n-1}{n} \sum_{i=1}^n \frac{1}{w_i}(\nabla_i f(x_k))^2. \tag{41}$$

From the optimality condition of Step 6 of Algorithm 1, and the fact that $X_i = \mathbb{R}$, we know that for all i the following holds,

$$\nabla_i f(x_k^{(1)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)}) = 0. \tag{42}$$

Therefore $\forall i$ we have

$$\begin{aligned}\frac{1}{w_i}(\nabla_i f(x_k))^2 &= \frac{1}{w_i}(\nabla_i f(x_k) - \nabla_i f(x_k^{(1)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)}))^2 \\ &\overset{(7)}{\leq} \frac{1}{w_i} L_i^2 (\tilde{x}^{(i)} - x_k^{(i)})^2 = \frac{1}{w_i^2} L_i^2 w_i (\tilde{x}^{(i)} - x_k^{(i)})^2.\end{aligned}$$

If we denote by $r^2 = \max_i \frac{L_i^2}{w_i^2}$, then we obtain from (41)

$$\begin{aligned}\mathbf{E}[(\|z_k\|_W^*)^2] &\overset{(40)}{\leq} \sum_{i=1}^n \left(\tfrac{1}{n} 2[(L_f^W)^2 + 1] + \tfrac{n-1}{n} r^2\right) w_i (x_k^{(i)} - \tilde{x}^{(i)})^2 \\ &= \left(\tfrac{1}{n} 2[(L_f^W)^2 + 1] + \tfrac{n-1}{n} r^2\right) \sum_{i=1}^n w_i (x_k^{(i)} - \tilde{x}^{(i)})^2 \\ &= \left(2[(L_f^W)^2 + 1] + (n-1) r^2\right) \tfrac{1}{n} \sum_{i=1}^n w_i (x_k^{(i)} - \tilde{x}^{(i)})^2 \\ &= \left(2[(L_f^W)^2 + 1] + (n-1) r^2\right) \mathbf{E}[\|x_k - x_{k+1}\|_W^2]\end{aligned}$$

and we can conclude that (15) holds with $\beta^2 = 2[(L_f^W)^2 + 1] + (n-1)r^2$.

Now, it remains to show (16). From (36) we know that

$$\nabla_i f((x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)},)^T)(\tilde{x}^{(i)} - x_k^{(i)}) \leq 0. \tag{43}$$

Therefore, from (17) with $\xi = x_k^{(i)}$ and $x = (x_k^{(1)}, \ldots, x_k^{(i-1)}, \tilde{x}^{(i)}, x_k^{(i+1)}, \ldots, x_k^{(n)})^T \overset{(37)}{=} x_{k+1}$, we have that

$$f(x_k) - f(x_{k+1}) \geq \gamma w_i |x_k^{(i)} - x_{k+1}^{(i)}|^2 + \nabla_i f(x_{k+1})(x_k^{(i)} - x_{k+1}^{(i)}) \overset{(43)}{\geq} \gamma w_i |x_k^{(i)} - x_{k+1}^{(i)}|^2. \tag{44}$$

Therefore

$$f(x_k) - f(x_{k+1}) \overset{(44)}{\geq} \gamma w_i |x_k^{(i)} - x_{k+1}^{(i)}|^2 = \gamma \|x_k - x_{k+1}\|_W^2.$$

and by taking expectation on both sides of the above, (16) follows with $\zeta = \gamma$.

Appendix B. Proof of Theorem 10

The proof is very similar to the proof of Theorem 7. Let us define an auxiliary vector $\tilde{x}$ in the same way as in (36). Then we can see that if coordinate i is chosen during iteration k in Algorithm 1 then (37) holds, and the optimality conditions for Step 6 of Algorithm 1 imply that (38) holds.

Note that x_{k+1} is a random variable which depends on i and x_k only. Therefore, we can define z_k such that i-th coordinate is given by (39). It is easy to verify that for z_k defined in (39), the condition (18) holds. Now, let us compute $(\|(z_k)_{[i]}\|_W^*)^2$. We have that

$$(\|(z_k)_{[i]}\|_W^*)^2 = \frac{1}{w_i}(z_k^{(i)})^2 \overset{(40)}{\leq} 2[(L_f^W)^2+1]w_i(x_k^{(i)}-\tilde{x}^{(i)})^2 \overset{(37)}{=} 2[(L_f^W)^2+1]\|x_k^{(i)}-x_{k+1}^{(i)}\|_W^2.$$

Therefore, we conclude that (19) holds with $\beta^2 = 2[(L_f^W)^2+1]$.

Now, it remains to show (20). Again from (36) we know that (43) holds. Therefore from (17) with $\xi = x_k^{(i)}$ and $x = (x_k^{(1)},\dots,x_k^{(i-1)},\tilde{x}^{(i)},x_k^{(i+1)},\dots,x_k^{(n)})^T \overset{(37)}{=} x_{k+1}$ we have (44). Therefore $f(x_k)-f(x_{k+1}) \overset{(44)}{\geq} \gamma w_i|x_k^{(i)}-x_{k+1}^{(i)}|^2 = \gamma\|x_k-x_{k+1}\|_W^2$, so (20) holds with $\zeta=\gamma$.

Appendix C. Proof of Theorem 12

This proof is based on the proof of Theorem 3.2 in Necoara (2015). We can write the optimality conditions for x_{k+1} from (14) and using the definition of a projection given in (9). We have that $\forall x \in X$, the following inequality holds

$$\left\langle W\left(x_{k+1}-x_k+\omega_k W^{-1}(\nabla f(x_k)-z_k)\right), x-x_{k+1}\right\rangle \geq 0. \tag{45}$$

Now, using the convexity of f we obtain that

$$\begin{aligned} f(x_{k+1})-f^* = f(x_{k+1})-f(\bar{x}_{k+1}) &\leq \langle \nabla f(x_{k+1}), x_{k+1}-\bar{x}_{k+1}\rangle \\ &= \langle \nabla f(x_{k+1})-\nabla f(x_k)+\nabla f(x_k), x_{k+1}-\bar{x}_{k+1}\rangle. \end{aligned} \tag{46}$$

Plugging $x=\bar{x}_{k+1}$ into (45) we obtain

$$\left\langle \tfrac{1}{\omega_k}W(x_{k+1}-x_k)-z_k, \bar{x}_{k+1}-x_{k+1}\right\rangle \geq \langle \nabla f(x_k), x_{k+1}-\bar{x}_{k+1}\rangle. \tag{47}$$

Plugging this into (46) gives us that

$$\begin{aligned} f(x_{k+1})-f(\bar{x}_{k+1}) &\overset{(46),(47)}{\leq} \left\langle \nabla f(x_{k+1})-\nabla f(x_k)-\tfrac{1}{\omega_k}W(x_{k+1}-x_k)+z_k, x_{k+1}-\bar{x}_{k+1}\right\rangle \\ &\overset{CS}{\leq} \|\nabla f(x_{k+1})-\nabla f(x_k)\|_W^*\|x_{k+1}-\bar{x}_{k+1}\|_W \\ &\quad + \left\langle -\tfrac{1}{\omega_k}W(x_{k+1}-x_k)+z_k, x_{k+1}-\bar{x}_{k+1}\right\rangle \\ &\overset{(8)}{\leq} L_f^W\|x_{k+1}-x_k\|_W\|x_{k+1}-\bar{x}_{k+1}\|_W \\ &\quad + \left\langle -\tfrac{1}{\bar{\omega}}W(x_{k+1}-x_k), x_{k+1}-\bar{x}_{k+1}\right\rangle + \langle z_k, x_{k+1}-\bar{x}_{k+1}\rangle \\ &\overset{CS}{\leq} L_f^W\|x_{k+1}-x_k\|_W\|x_{k+1}-\bar{x}_{k+1}\|_W] \\ &\quad + \tfrac{1}{\bar{\omega}}\|W(x_{k+1}-x_k)\|_W^*\|x_{k+1}-\bar{x}_{k+1}\|_W + \|z_k\|_W^*\|x_{k+1}-\bar{x}_{k+1}\|_W \\ &= \left((L_f^W+\tfrac{1}{\bar{\omega}})\|x_{k+1}-x_k\|_W + \|z_k\|_W^*\right)\|x_{k+1}-\bar{x}_{k+1}\|_W \\ &\overset{(5)}{\leq} \left((L_f^W+\tfrac{1}{\bar{\omega}})\|x_{k+1}-x_k\|_W + \|z_k\|_W^*\right)\sqrt{\frac{1}{\kappa_f}(f(x_{k+1})-f(\bar{x}_{k+1}))}. \end{aligned} \tag{48}$$

Therefore, we can conclude that

$$f(x_{k+1}) - f^* \overset{(48)}{\leq} \frac{1}{\kappa_f}\left((L_f^W + \tfrac{1}{\bar{\omega}})\|x_{k+1} - x_k\|_W + \|z_k\|_W^*\right)^2. \tag{49}$$

Taking the expectation of (49) with respect to the random vector z_k, we obtain

$$\begin{aligned}
\mathbf{E}[f(x_{k+1}) - f(\bar{x}_{k+1})] &\overset{(49)}{\leq} \frac{1}{\kappa_f}\mathbf{E}\left[\left((L_f^W + \tfrac{1}{\bar{\omega}})\|x_{k+1} - x_k\|_W + \|z_k\|_W^*\right)^2\right] \\
&\leq \frac{2}{\kappa_f}\left((L_f^W + \tfrac{1}{\bar{\omega}})^2\mathbf{E}[\|x_{k+1} - x_k\|_W^2] + \mathbf{E}[(\|z_k\|_W^*)^2]\right) \\
&\overset{(15)}{\leq} \frac{2}{\kappa_f}\left((L_f^W + \tfrac{1}{\bar{\omega}})^2 + \beta^2\right)\mathbf{E}[\|x_k - x_{k+1}\|_W^2] \\
&\overset{(16)}{\leq} \frac{2}{\kappa_f}\left((L_f^W + \tfrac{1}{\bar{\omega}})^2 + \beta^2\right)\frac{1}{\zeta}\left(f(x_k) - \mathbf{E}[f(x_{k+1})]\right) \\
&= \underbrace{\frac{2}{\kappa_f}\left((L_f^W + \tfrac{1}{\bar{\omega}})^2 + \beta^2\right)\frac{1}{\zeta}}_{c}\Big(f(x_k) - f(\bar{x}_k) \\
&\quad + \mathbf{E}[f(\bar{x}_{k+1})] - \mathbf{E}[f(x_{k+1})]\Big).
\end{aligned} \tag{50}$$

Finally, from (50) we obtain that

$$\mathbf{E}[f(x_{k+1}) - f^*] = \mathbf{E}[f(x_{k+1}) - f(\bar{x}_{k+1})] \leq \frac{c}{1+c}\left(f(x_k) - f(\bar{x}_{k+1})\right) = \frac{c}{1+c}\left(f(x_k) - f^*\right),$$

and the result follows.

Appendix D. Proof of Theorem 13 if $z_k = 0$

Let us define an auxiliary vector $\tilde{x}$ such that

$$\tilde{x}^{(i)} = \mathbf{Proj}_X^W\left(x_k - \omega_k W^{-1}(\nabla f(x_k) - z_k)_{[\mathcal{I}]}\right)_{[\mathcal{I}]}. \tag{51}$$

Then we can see that if coordinates $\mathcal{I}$ is chosen during iteration k in Algorithm 1 then

$$x_{k+1}^{(j)} = \begin{cases} x_k^{(j)}, & \text{if } \ j \notin \mathcal{I}, \\ \tilde{x}^{(i)}, & \text{otherwise.} \end{cases} \tag{52}$$

Therefore, let us estimate the expected value of f at a random point x_{k+1}, where the expectation is taken with respect to the selection of coordinates $\mathcal{I}$ at iteration k. Let

$h \in \mathbb{R}^n$. Then if $\frac{1}{\omega_k} \geq \max_i \frac{L_i}{w_i}$ we have

$$\begin{aligned}
\mathbf{E}[f(x_k + h_{[\mathcal{I}]})] &\overset{(7)}{\leq} f(x_k) + \mathbf{E}\left[\left\langle \nabla f(x_k), h_{[\mathcal{I}]}\right\rangle + \frac{L_{\mathcal{I}}}{2w_{\mathcal{I}}}\|h_{[\mathcal{I}]}\|_W^2\right] \\
&\leq f(x_k) + \mathbf{E}\left[\left\langle \nabla f(x_k), h_{[\mathcal{I}]}\right\rangle + \frac{1}{2\omega_k}\|h_{[\mathcal{I}]}\|_W^2\right] \\
&\overset{(52)}{=} f(x_k) + \frac{\tau}{n}\left(\langle \nabla f(x_k), h\rangle + \frac{1}{2\omega_k}\|h\|_W^2\right) \\
&= \frac{n-\tau}{n} f(x_k) + \frac{\tau}{n}\left(\underbrace{f(x_k) + \langle \nabla f(x_k) - z_k, h\rangle + \frac{1}{2\omega_k}\|h\|_W^2}_{\mathcal{H}(h;x_k,z_k)} + \langle z_k, h\rangle\right).
\end{aligned} \tag{53}$$

Now, observe that

$$\begin{aligned}
\tilde{x} &= x_k + \arg\min_{h: x+x_k \in X} \mathcal{H}(h; x_k, z_k) \\
&= x_k + \arg\min_{h \in \mathbb{R}^n} \{\mathcal{H}(h; x_k, z_k) + \Phi_X(x + x_k)\} =: x_k + \hat{h},
\end{aligned} \tag{54}$$

where $\Phi_X(x)$ is the indicator function for the set X, i.e.

$$\Phi_X(x) = \begin{cases} 0, & \text{if } x \in X, \\ +\infty, & \text{otherwise.} \end{cases} \tag{55}$$

From the first order optimality conditions of (54) we have

$$\nabla f(x_k) - z_k + \frac{1}{\omega_k} W\hat{h} + s = 0, \tag{56}$$

where $s \in \partial\Phi(x_k + \hat{h})$. We can define a composite gradient mapping Lu and Xiao (2013); Nesterov (2013); Tappenden et al. (2015) as

$$g := -\frac{1}{\omega_k} W\hat{h}. \tag{57}$$

Therefore, we can observe that

$$-\nabla f(x_k) + z_k + g \overset{(56)}{\in} \partial\Phi(x_k + \hat{h}). \tag{58}$$

It is also easy to show that

$$\|\hat{h}\|_W^2 = \|\omega_k W^{-1} g\|_W^2 = \omega_k^2 (\|g\|_W^*)^2 \tag{59}$$

and

$$\left\langle g, \hat{h}\right\rangle = -\frac{1}{\omega_k}\|\hat{h}\|_W^2 \overset{(59)}{=} -\omega_k (\|g\|_W^*)^2. \tag{60}$$

Finally note that for any $y \in X$ we have

$$\begin{aligned}\|x_k + \hat{h} - y\|_W^2 &= \|x_k - y\|_W^2 + 2\omega_k \langle g, y - x_k\rangle + \|\hat{h}\|_W^2 \\ &\overset{(59)}{=} \|x_k - y\|_W^2 + 2\omega_k \langle g, y - x_k\rangle + \omega_k^2(\|g\|_W^*)^2. \end{aligned} \tag{61}$$

Now, we are ready to bound $\mathcal{H}(h; x_k, z_k) + \Phi(x + h)$ for $h = \hat{h}$. We have

$$\begin{aligned}
&\mathcal{H}(\hat{h}; x_k, z_k) + \Phi(x_k + \hat{h}) \\
&= f(x_k) + \left\langle \nabla f(x_k) - z_k, \hat{h}\right\rangle + \frac{1}{2\omega_k}\|\hat{h}\|_W^2 + \Phi(x_k + \hat{h}) \\
&\overset{(58)}{\leq} f(y) + \langle \nabla f(x_k), x_k - y\rangle + \left\langle \nabla f(x_k) - z_k, \hat{h}\right\rangle + \frac{1}{2\omega_k}\|\hat{h}\|_W^2 \\
&\quad + \Phi(y) + \left\langle -\nabla f(x_k) + z_k + g, x_k + \hat{h} - y\right\rangle \\
&= f(y) + \Phi(y) + \frac{1}{2\omega_k}\|\hat{h}\|_W^2 + \langle g, x_k - y\rangle + \langle z_k, x_k - y\rangle + \left\langle g, \hat{h}\right\rangle \\
&\overset{(60),(59)}{=} f(y) + \Phi(y) + \frac{1}{2}\omega_k(\|g\|_W^*)^2 + \langle g, x_k - y\rangle + \langle z_k, x_k - y\rangle - \omega_k(\|g\|_W^*)^2 \\
&= f(y) + \Phi(y) - \frac{1}{2}\omega_k(\|g\|_W^*)^2 + \langle g, x_k - y\rangle + \langle z_k, x_k - y\rangle \\
&\overset{(61)}{=} f(y) + \Phi(y) - \frac{1}{2\omega_k}\left(\|x_k + \hat{h} - y\|_W^2 - \|x_k - y\|_W^2\right) + \langle z_k, x_k - y\rangle \\
&\overset{(52),(63)}{=} f(y) + \Phi(y) - \frac{1}{2\omega_k}\frac{n}{\tau}\left(\mathbf{E}[\|x_{k+1} - y\|_W^2] - \|x_k - y\|_W^2\right) + \langle z_k, x_k - y\rangle,
\end{aligned}$$

where in the last step, we use

$$n\mathbf{E}[\|x_{k+1} - y\|_W^2] = \tau\|x_k + \hat{h} - y\|_W^2 + (n - \tau)\|x_k - y\|_W^2.$$

Now, from (53) we conclude that $\forall y$ we have

$$\begin{aligned}\mathbf{E}[f(x_{k+1})] \leq \frac{n-\tau}{n}f(x_k) + \frac{\tau}{n}\Bigg(&f(y) + \Phi(y) - \frac{n}{2\omega_k\tau}\mathbf{E}[\|x_{k+1} - y\|_W^2] \\ &+ \frac{n}{2\omega_k\tau}\|x_k - y\|_W^2 + \left\langle z_k, x_k + \hat{h} - y\right\rangle\Bigg),\end{aligned}$$

which can be equivalently written as

$$\begin{aligned}\mathbf{E}\left[f(x_{k+1}) + \frac{1}{2\omega_k}\|x_{k+1} - y\|_W^2\right] \leq &f(x_k) + \frac{1}{2\omega_k}\|x_k - y\|_W^2 \\ &- \frac{\tau}{n}(f(x_k) - f(y) - \Phi(y)) + \frac{\tau}{n}\left\langle z_k, x_k + \hat{h} - y\right\rangle.\end{aligned}$$

If we choose $y = \bar{x}_k$ then the latter inequality reads as follows

$$\begin{aligned}\mathbf{E}\left[f(x_{k+1}) + \frac{1}{2\omega_k}\|x_{k+1} - \bar{x}_k\|_W^2\right] \leq &f(x_k) + \frac{1}{2\omega_k}\|x_k - \bar{x}_k\|_W^2 \\ &- \frac{\tau}{n}(f(x_k) - f^*) + \frac{\tau}{n}\left\langle z_k, x_k + \hat{h} - \bar{x}_k\right\rangle.\end{aligned}$$

From the definition of $\bar{x}$ we obtain that $\|x_{k+1} - \bar{x}_{k+1}\|_W \leq \|x_{k+1} - \bar{x}_k\|_W$ and therefore

$$\begin{aligned}\mathbf{E}\left[f(x_{k+1}) - f^* + \frac{1}{2\omega_k}\|x_{k+1} - \bar{x}_{k+1}\|_W^2\right] \leq & (1 - \tfrac{\tau}{n})(f(x_k) - f^*) \\ & + \frac{1}{2\omega_k}\|x_k - \bar{x}_k\|_W^2 + \frac{\tau}{n}\left\langle z_k, x_k + \hat{h} - \bar{x}_k\right\rangle.\end{aligned}$$

Let us assume that $\forall k : z_k = 0$. Then let us define $c = \frac{2\tau\bar{\omega}}{n(2\bar{\omega}+1)} \in (0,1)$,

$$\mathbf{E}\left[f(x_{k+1}) - f^* + \frac{1}{2\omega_k}\|x_{k+1} - \bar{x}_{k+1}\|_W^2\right] \leq (1-c)\left(f(x_k) - f^* + \frac{1}{2\bar{\omega}}\|x_k - \bar{x}_k\|_W^2\right). \quad (62)$$

Therefore,

$$\begin{aligned}\mathbf{E}[f(x_k) - f^*] &\leq \mathbf{E}\left[f(x_k) - f^* + \frac{1}{2\omega_k}\|x_k - \bar{x}_k\|_W^2\right] \\ &\overset{(62)}{\leq} (1-c)^k\left(f(x_0) - f^* + \frac{1}{2\bar{\omega}}\|x_0 - \bar{x}_0\|_W^2\right).\end{aligned}$$

Appendix E. Proof of Theorem 13 if $z_k \neq 0$

The proof follows similar arguments to the proof of Theorem 13 when $z_k = 0$. Let us define an auxiliary vector $\tilde{x}$ in the same way as in (51). Then we can see that if coordinates $\mathcal{I}$ is chosen during iteration k in Algorithm 1 then (52) holds. Therefore, let us estimate the expected value of f at a random point x_{k+1}, where the expectation is taken with respect to the selection of coordinate i at iteration k. Let $h \in \mathbb{R}^n$. Then if $\frac{1}{\omega_k} \geq \max_i \frac{L_i}{w_i}$ we have that (53) holds. Now, observe that

$$\begin{aligned}\tilde{x} &= x_k + \arg\min_{h: x + x_k \in X} \mathcal{H}(h; x_k, z_k) \\ &= x_k + \arg\min_{h \in \mathbb{R}^n}\{\mathcal{H}(h; x_k, z_k) + \Phi_X(h + x_k)\} =: x_k + \hat{h}, \end{aligned} \quad (63)$$

where $\Phi_X(x)$ is indicator function for set X, (55). Now, we have

$$\begin{aligned}\mathcal{H}(\hat{h}; x_k, z_k) &= \min_{h \in \mathbb{R}^n}\left\{f(x_k) + \langle \nabla f(x_k) - z_k, h\rangle + \frac{1}{2\omega_k}\|h\|_W^2 + \Phi_X(h + x_k)\right\} \\ &= \min_{y \in \mathbb{R}^n}\left\{f(x_k) + \langle \nabla f(x_k) - z_k, y - x_k\rangle + \frac{1}{2\omega_k}\|y - x_k\|_W^2 + \Phi_X(y)\right\} \\ &\leq \min_{\lambda \in [0,1]}\Bigg\{f(\lambda\bar{x}_k + (1-\lambda)x_k) + \langle -z_k, \lambda(\bar{x}_k - x_k)\rangle \\ &\qquad + \frac{1}{2\omega_k}\|\lambda(\bar{x}_k - x_k)\|_W^2 + \Phi_X(\lambda(\bar{x}_k - x_k) + x_k)\Bigg\} \\ &\leq \min_{\lambda \in [0,1]}\left\{\lambda f(\bar{x}_k) + (1-\lambda)f(x_k) + \lambda\|z_k\|_W^*\|\bar{x}_k - x_k\|_W + \frac{\lambda^2}{2\omega_k}\|\bar{x}_k - x_k\|_W^2\right\}.\end{aligned}$$

Note that from (52) and (63) we have

$$\|\hat{h}\|_W^2 = \sum_{i=1}^n \|\hat{h}_{[\mathcal{I}]}\|_W^2 = \frac{n}{\tau}\mathbf{E}[\|x_{k+1} - x_k\|_W^2] \overset{(20)}{\leq} \frac{n}{\zeta\tau}\mathbf{E}[f(x_k) - f(x_{k+1})]. \quad (64)$$

Therefore, we conclude that

$$\begin{aligned}\mathbf{E}[f(x_{k+1})-f^*] &\overset{(53),(19)}{\leq} \min_{\lambda\in[0,1]}\Big\{f(x_k)-f^*+\frac{\tau}{n}\Big(\lambda(f(\bar{x}_k)-f(x_k))+\lambda\|z_k\|_W^*\|\bar{x}_k-x_k\|_W\\ &\quad+\frac{\lambda^2}{2\omega_k}\|\bar{x}_k-x_k\|_W^2+\|z_k\|_W^*\|\hat{h}\|_W\Big)\Big\}\\ &\overset{(5)}{\leq} \min_{\lambda\in[0,1]}\Big\{f(x_k)-f^*+\tfrac{\tau}{n}\Big(-\lambda(f(x_k)-f^*)+\lambda\|z_k\|_W^*\|\bar{x}_k-x_k\|_W\\ &\quad+\frac{\lambda^2}{2\omega_k\kappa_f}(f(x_k)-f^*)+\|z_k\|_W^*\|\hat{h}\|_W\Big)\Big\}.\end{aligned}$$

Now, let us denote by $\xi_k=f(x_k)-f^*$. Notice that

$$(\|z_k\|_W^*)^2=\sum_{i=1}^n(\|(z_k)_{[\mathcal{I}]}\|_W^*)^2\overset{(19),(20)}{\leq} n\frac{\beta^2}{\zeta}(\xi_k-\mathbf{E}[\xi_{k+1}]) \tag{65}$$

where the expectation is with respect to the random choice i during the k-th iteration. Therefore we have

$$\begin{aligned}\mathbf{E}[\xi_{k+1}] &\leq \min_{\lambda\in[0,1]}\Big\{\xi_k+\tfrac{\tau}{n}\Big(-\lambda\xi_k+\lambda\|z_k\|_W^*\|\bar{x}_k-x_k\|_W+\frac{\lambda^2}{2\omega_k\kappa_f}\xi_k+\|z_k\|_W^*\|\hat{h}\|_W\Big)\Big\}\\ &\overset{(65),(64)}{\leq} \min_{\lambda\in[0,1]}\Big\{\xi_k+\tfrac{\tau}{n}\Big(-\lambda\xi_k+\lambda\|z_k\|_W^*\|\bar{x}_k-x_k\|_W\\ &\quad+\frac{\lambda^2}{2\omega_k\kappa_f}\xi_k+\frac{n\beta}{\zeta\sqrt{\tau}}(\xi_k-\mathbf{E}[\xi_{k+1}])\Big)\Big\}\end{aligned}$$

which is equivalent to

$$\begin{aligned}(1+\tfrac{\sqrt{\tau}\beta}{\zeta})\mathbf{E}[\xi_{k+1}] &\leq (1+\tfrac{\sqrt{\tau}\beta}{\zeta})\xi_k+\min_{\lambda\in[0,1]}\{-\tfrac{\tau}{n}\lambda\xi_k+\tfrac{\tau}{n}\lambda\|z_k\|_W^*\|\bar{x}_k-x_k\|_W+\tfrac{\tau}{n}\tfrac{\lambda^2}{2\omega_k\kappa_f}\xi_k\}\\ &\overset{(65),(5)}{\leq} (1+\tfrac{\sqrt{\tau}\beta}{\zeta})\xi_k+\min_{\lambda\in[0,1]}\{-\tfrac{\tau}{n}\lambda\xi_k+\tfrac{\tau}{n}\lambda\sqrt{n\tfrac{\beta^2}{\zeta}(\xi_k-\mathbf{E}[\xi_{k+1}])}\sqrt{\tfrac{1}{\kappa_f}\xi_k}+\tfrac{\tau}{n}\tfrac{\lambda^2}{2\omega_k\kappa_f}\xi_k\}.\end{aligned}$$

Using the fact that $\forall a,b\in\mathbb{R}_+$ we have $\sqrt{ab}\leq\frac{1}{2}a+\frac{1}{2}b$ we obtain that

$$\begin{aligned}&(1+\tfrac{\sqrt{\tau}\beta}{\zeta})\mathbf{E}[\xi_{k+1}]\\ \leq&(1+\tfrac{\sqrt{\tau}\beta}{\zeta})\xi_k+\min_{\lambda\in[0,1]}\{-\tfrac{\tau}{n}\lambda\xi_k+\sqrt{\tfrac{\beta^2}{\zeta}(\xi_k-\mathbf{E}[\xi_{k+1}])}\sqrt{\tfrac{\lambda^2}{n}\tfrac{\tau^2}{\kappa_f}\xi_k}+\tfrac{\tau}{n}\tfrac{\lambda^2}{2\omega_k\kappa_f}\xi_k\}\\ \leq&(1+\tfrac{\sqrt{\tau}\beta}{\zeta})\xi_k+\min_{\lambda\in[0,1]}\{-\tfrac{\tau}{n}\lambda\xi_k+\frac{\beta^2}{2\zeta}(\xi_k-\mathbf{E}[\xi_{k+1}])+\frac{1}{2}\frac{\lambda^2}{n}\frac{\tau^2}{\kappa_f}\xi_k+\tfrac{\tau}{n}\tfrac{\lambda^2}{2\omega_k\kappa_f}\xi_k\}.\end{aligned}$$

Therefore, we obtain

$$(1+\tfrac{\sqrt{\tau}\beta}{\zeta}+\tfrac{\beta^2}{2\zeta})\mathbf{E}[\xi_{k+1}]\leq(1+\tfrac{\sqrt{\tau}\beta}{\zeta}+\tfrac{\beta^2}{2\zeta})\xi_k+\frac{\tau}{n\bar{\omega}\kappa_f}\min_{\lambda\in[0,1]}\{-\lambda\bar{\omega}\kappa_f+\frac{\lambda^2}{2}(1+\bar{\omega}\tau)\}\xi_k. \tag{66}$$

The optimal λ^* that minimizes the above expression is

$$\lambda^* = \min\left\{1, \frac{\bar{\omega}\kappa_f}{\bar{\omega}\tau + 1}\right\}.$$

Consider now two cases:

- $\lambda^* < 1$. In this case

$$-\lambda^*\bar{\omega}\kappa_f + \frac{(\lambda^*)^2}{2}(1+\bar{\omega}\tau) = -\frac{1}{2}\frac{(\bar{\omega}\kappa_f)^2}{\bar{\omega}\tau+1}.$$

  Combining this with (66) gives

$$(1 + \tfrac{\sqrt{\tau}\beta}{\zeta} + \tfrac{\beta^2}{2\zeta})\mathbf{E}[\xi_{k+1}] \leq (1 + \tfrac{\sqrt{\tau}\beta}{\zeta} + \tfrac{\beta^2}{2\zeta} - \frac{1}{2n}\frac{\bar{\omega}\kappa_f}{\bar{\omega}\tau+1})\xi_k,$$

  which is equivalent to

$$\mathbf{E}[\xi_{k+1}] \leq \left(1 - \frac{1}{2n}\frac{2\bar{\omega}\kappa_f\zeta}{(\bar{\omega}\tau+1)(2\zeta + 2\beta\sqrt{\tau}+\beta)}\right)\xi_k.$$

- $\lambda^* = 1$. In this case $\frac{\bar{\omega}\kappa_f}{\bar{\omega}\tau+1} \geq 1$ and hence

$$-\lambda^*\bar{\omega}\kappa_f t + \frac{(\lambda^*)^2}{2}(1+\bar{\omega}\tau) = -\bar{\omega}\kappa_f + \frac{1}{2}(1+\bar{\omega}\tau) \leq -\bar{\omega}\kappa_f + \frac{1}{2}\bar{\omega}\kappa_f = -\frac{1}{2}\bar{\omega}\kappa_f.$$

  Therefore, from (66) we can conclude that

$$\mathbf{E}[\xi_{k+1}] \leq \left(1 - \frac{\zeta\tau}{n(2\zeta + 2\beta\sqrt{\tau} + 1 + \beta^2)}\right)\xi_k.$$

References

Silvia Bonettini. Inexact block coordinate descent methods with application to non-negative matrix factorization. *IMA Journal of Numerical Analysis*, 31(4):1431–1452, 2011.

Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. Coordinate descent method for large-scale l2-loss linear support vector machines. *The Journal of Machine Learning Research*, 9:1369–1398, 2008.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and Sellamanickam Sundararajan. A dual coordinate descent method for large-scale linear svm. In *International Conference on Machine Learning (ICML)*, 2008.

Xiaoqin Hua and Nobuo Yamashita. An inexact coordinate descent method for the weighted l1-regularized convex optimization problem. Technical report, Technical report, School of Mathematics and Physics, Kyoto University, Kyoto 606-8501, Japan, 2012.

Wu Li. The sharp Lipschitz constants for feasible and optimal solutions of a perturbed linear program. *Linear algebra and its applications*, 187:15–40, 1993.

Ji Liu and Stephen J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):35117376, 2015.

Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, pages 1–28, 2013.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Ion Necoara. Linear convergence of first order methods under weak nondegeneracy assumptions for convex programming. *arXiv:1504.06298*, 2015.

Ion Necoara and Dragos Clipici. Parallel coordinate descent methods for composite minimization. *SIAM Journal on Optimization*, 26(1):19717226, 2016.

Ion Necoara and Valentin Nedelcu. Distributed dual gradient methods and error bound conditions. *arXiv:1401.4398*, 2014a.

Ion Necoara and Valentin Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *Automatic Control, IEEE Transactions on*, 59(5):1232–1243, 2014b.

Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144 (1-2):1–38, 2014.

Stephen M. Robinson. Bounds for error in the solution set of a perturbed linear program. *Linear Algebra and its applications*, 6:69–81, 1973.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

Ruoyu Sun and Yinyu Ye. Worst-case complexity of cyclic coordinate descent: $o(n^2)$ gap with randomized version. *arXiv:1604.07130*, 2016.

Martin Takáč, Peter Richtárik, and Nathan Srebro. Distributed mini-batch SDCA. *arXiv:1507.08322*, 2015.

Martin Takáč, Avleen Singh Bijral, Peter Richtárik, and Nathan Srebro. Mini-batch primal and dual methods for SVMs. *International Conference on Machine Learning (ICML)*, 2013.

Rachael Tappenden, Martin Takáč, and Peter Richtárik. On the complexity of parallel coordinate descent. *arXiv:1503.03033*, 2015.

Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact coordinate descent: complexity and preconditioning. *Journal of Optimization Theory and Applications*, 170(1): 14417176, 2016.

Po-Wei Wang and Chih-Jen Lin. Iteration complexity of feasible descent methods for convex optimization. *The Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

H. Zhang and W. Yin. Gradient methods for convex minimization: better rates under weaker conditions. Technical report, CAM Report 13-17, UCLA, 2013.