DOI: 10.1515/auom-2015-0052
An. Şt. Univ. Ovidius Constanţa VERSITA **Vol. 23(3),2015, 129–149**

Randomized Sparse Block Kaczmarz as Randomized Dual Block-Coordinate Descent

Stefania Petra

Abstract

We show that the Sparse Kaczmarz method is a particular instance of the coordinate gradient method applied to an unconstrained dual problem corresponding to a regularized ℓ_1-minimization problem subject to linear constraints. Based on this observation and recent theoretical work concerning the convergence analysis and corresponding convergence rates for the randomized block coordinate gradient descent method, we derive block versions and consider randomized ordering of blocks of equations. Convergence in expectation is thus obtained as a byproduct. By smoothing the ℓ_1-objective we obtain a strongly convex dual which opens the way to various acceleration schemes.

1 Introduction

Overview. We consider an *underdeterimed* matrix $A \in \mathbb{R}^{m\times n}$, with $m < n$, $b \in \mathbb{R}^m$ and wish to solve the problem

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \quad \text{s.t.} \quad Ax = b. \tag{1}$$

Key Words: Randomized block-coordinate descent, underdetermined system of linear equations, compressed sensing, sparse Kaczmarz solver, linearized Bregman method, convex duality, convex programming, limited-data tomography.

2010 Mathematics Subject Classification: Primary 65K10, 90C25; Secondary 65F22, 68U10.

Received: January, 2015.
Revised: February, 2015.
Accepted: March, 2015.

The minimal ℓ_1-norm solution is relevant for many applications and also favors *sparse* solutions, provided some assumption on A hold, see [8]. Here we consider a *regularized* version of (1)

$$\min_{x\in\mathbb{R}^n} \lambda\|x\|_1 + \frac{1}{2}\|x\|^2 \quad \text{s.t.} \quad Ax = b. \tag{2}$$

We note that for a large but finite parameter $\lambda > 0$, the solution of (2) gives a minimal ℓ_1-norm solution of $Ax = b$ [9]. The advantage of (2) is the strongly convex objective function. In [11] this property was used to design an iterative method employing Bregman projections with respect to $f(x) = \lambda\|x\|_1 + \frac{1}{2}\|x\|^2$ on subsets of equations in $Ax = b$. The convergence of the method was shown in the framework of the *split feasibilty problems* [7, 5]. We will follow a different approach and consider a block coordinate gradient descent method applied to the dual of (2). This dual problem is unconstrained and differentiable due to the - in particular - strict convexity of f. It will turn out that the two methods are equivalent. For example, the iterative Bregman projection method previously mentioned, gives the *Sparse Kaczmarz* method when Bregman projections are performed with respect to f on each single hyperplane defined by each row of A. On the other had, it is a coordinate gradient descent method applied on the dual problem of (2).

Contribution and Organization. Beyond the interpretation of the Sparse Kaczmarz method as a coordinate descent method on the dual problem, we consider block updates by performing block coordinate gradient descent steps on the dual problem. This will lead to the Block Sparse Kaczmarz method. The block order is left to a stochastic process in order to exploit recent results in the field of the randomized block gradient descent methods. By exploiting this parallelism between primal and dual updates, we obtain convergence as a byproduct and *convergence rates* which have not been available so far. We also consider smoothing techniques for acceleration purposes.

We introduce the Randomized Block Sparse Kaczmarz Algorithm in Section 2. In Section 3 we derive the dual problem of (2) and its smoothed version and establish the connection between primal and dual variables. We review the literature on the randomized block gradient descent method for smooth unconstrained optimization in Section 4. Section 5 presents the link between the two methods. We conclude with Section 6 where we present numerical examples on sparse tomographic recovery from few projections.

Notation. For $m \in \mathbb{N}$, we denote $[m] = \{1, 2, \ldots, m\}$. For some matrix A and a vector z, A_J denotes the submatrix of rows indexed by J, and

z_J the corresponding subvector. A_i will denote the ith row of A. $\mathcal{R}(A)$ denotes the range of A. Vectors are column vectors and indexed by superscripts. $A^\top$ denotes the transposed of A. $\langle x, z\rangle$ denotes the standard scalar product in $\mathbb{R}^n$ and $\|x\|_2 = \sqrt{\langle x, x\rangle}$, while $\|x\|_1 = \sum_{i\in[n]} |x_i|$. With $\langle x, y\rangle_w$ we denote the weighted inner product $\sum_{i\in[n]} w_i x_i y_i$ with $x, y \in \mathbb{R}^n$, $w \in \mathbb{R}^n_{++}$. The indicator function of a set C is denoted by $\delta_C(x) := \begin{cases} 0, & \text{if } x \in C \\ +\infty, & \text{if } x \notin C \end{cases}$. $\sigma_C(x) := \sup_{y\in C}\langle y, x\rangle$ denotes the support function of a nonempty set C. $\partial f(x)$ is the subdifferential and $\nabla f(x)$ the gradient of f at x and $\operatorname{int} C$ and $\operatorname{rint} C$ denote the interior and the relative interior of a set C. By f^* we denote the conjugate function of f. We refer to the Appendix and [19] for related properties. Δ_n denotes the probability simplex $\Delta_n = \{x \in \mathbb{R}^n_+ \colon \langle \mathbb{1}, x\rangle = 1\}$. $\mathbb{E}_p[\cdot]$ denotes the expectation with respect to the distribution p; the subscript p is omitted if p is clear from the context.

2 A Randomized Block Sparse Kaczmarz Method

Consider $A \in \mathbb{R}^{m\times n}$ and let $\cup_{i\in[c]} S_i = [m]$ be a collection of subsets – not necessarily a partition – covering $[m]$. Further, consider the *consistent* – possibly *underdetermined* – system of equations

$$Ax = b, \tag{3}$$

which can be rewritten as

$$\hat{A}x = \hat{b}, \quad \hat{A} := \begin{pmatrix} A_{S_1} \\ \vdots \\ A_{S_c} \end{pmatrix} \in \mathbb{R}^{\hat{m}\times\hat{n}}, \qquad \hat{b} := \begin{pmatrix} b_{S_1} \\ \vdots \\ b_{S_c} \end{pmatrix} \in \mathbb{R}^{\hat{m}}, \tag{4}$$

without changing the solution set, provided $b \in \mathcal{R}(A)$, which we will further assume to hold.

We consider the problem (2). A simple method for the solution of (2), called *linearized Bregman method* [22, 6], is

$$z^{(k+1)} = z^{(k)} - t_k A^\top (Ax^{(k)} - b), \tag{5}$$

$$x^{(k+1)} = S_\lambda(z^{(k+1)}), \tag{6}$$

initialized with $z^{(0)} = x^{(0)} = 0$. $S_\lambda(x) = \operatorname{sign}(x)\max(|x| - \lambda, 0)$ denotes the soft shrinkage operator. According to [11], not only the constant stepsize $t_k = \frac{1}{\|A\|_2^2}$, like chosen in [6], leads to convergence, but also the *dynamic* stepsize choices (depending on the iterate $x^{(k)}$), and *exact* stepsizes obtained from a univariate minimization scheme. Here, however, we concentrate on constant stepsize choices only.

In the general framework of *split feasibilty problems* [7, 5] and Bregman projections, convergence of the *Sparse Kaczmarz method*

$$z^{(k+1)} = z^{(k)} - t_k \left(\langle A_{i_k}, x^{(k)} \rangle - b_{i_k} \right) A_{i_k}^\top, \tag{7}$$

$$x^{(k+1)} = S_\lambda(z^{(k+1)}), \tag{8}$$

is shown [11], with $z^0 = x^0 = 0$ and i_k an appropriate *control sequence* (e.g. *cyclic*) for row selection. Here a *single row* of A is used, thus $S_i = \{i_k\}$.

In view of [11, Cor. 2.9], one could also work *block-wise* and consider groups of equations $A_{S_i}x = b_{S_i}$ to obtain the *Block Sparse Kaczmarz* scheme

$$z^{(k+1)} = z^{(k)} - t_k(A_{S_{i_k}})^\top(A_{S_{i_k}}x^{(k)} - b_{S_{i_k}}) \tag{9}$$

$$x^{(k+1)} = S_\lambda(z^{(k+1)}). \tag{10}$$

Convergence of the scheme would follow by choosing block $A_{S_{i_k}}$ from (4) in (any) cyclic order, see the considerations in [11]. Motivated by these results and by the fact that *convergence rates* for the above methods are not known, we introduce the following *Randomized Block Sparse Kaczmarz* scheme, Alg. 1. We will address next how to choose the probability vector $p \in \Delta_c$, and thus the random choice of blocks of A, along with the stepsize t_k in order to obtain a convergent scheme.

Algorithm 1: Randomized Block Sparse Kaczmarz (RBSK)

```
Input: Starting vectors $x^{(0)}$ and $z^{(0)}$, covering of $[m] = \cup_{i=1}^c S_i$, with
       $|S_i| = m_i$, $i \in [c]$, probability vector $p \in \Delta_c$
for $k = 1, 2, \dots$ do
    Sample $i_k \in [c]$ due $i_k \sim p$
    Update $z^{(k+1)} = z^{(k)} - t_k(A_{S_{i_k}})^\top(A_{S_{i_k}}x^{(k)} - b_{S_{i_k}})$
    Update $x^{(k+1)} = S_\lambda(z^{(k)})$
```

3 Primal and Dual Problems

In this Section we derive the dual problem of (2) and the relation between primal and dual variables. This will be the backbone of the observed equivalence. Moreover, we consider a smoothed version of (2) along with its dual.

3.1 Minimal ℓ_1-Norm Solution via the Dual

Primal Problem Consider the primal problem (2). We write (2) in the form (49a)

$$\min \varphi(x), \quad \varphi(x) := \langle 0, x\rangle + \underbrace{\lambda\|x\|_1 + \frac{1}{2}\|x\|^2}_{:=f(x)} + \delta_{\{0\}}(b - Ax). \tag{11}$$

Denoting $g := \delta_{\{0\}}$, we get $g^* \equiv 0$. On the other hand, we have $f^*(y) = \frac{1}{2}\|S_\lambda(y)\|^2$. Indeed, computing the Fenchel conjugate of f we obtain

$$\begin{aligned} f^*(y) &= \sup_x \{y^\top x - \lambda\|x\|_1 - \frac{1}{2}\|x\|^2\} = \sum_{i\in[n]} \max_{x_i}\{y_i x_i - \lambda|x_i| - \frac{1}{2}x_i^2\} \\ &= \sum_{i: y_i > \lambda} (y_i(y_i - \lambda) - \lambda(y_i - \lambda) - \frac{1}{2}(y_i - \lambda)^2) + \sum_{i:|y_i|\le\lambda} 0 \\ &+ \sum_{i: y_i < -\lambda} (y_i(y_i + \lambda) + \lambda(y_i + \lambda) - \frac{1}{2}(y_i + \lambda)^2) \\ &= \frac{1}{2}\|y - \Pi_{[-\lambda,\lambda]^n}(y)\|^2 = \frac{1}{2}\|S_\lambda(y)\|^2 \ . \end{aligned}$$

Dual Problem Now (49b) in A.2 (Fenchel duality formula in the Appendix) gives the dual problem

$$\inf \psi(y), \quad \psi(y) := -\langle b, y\rangle + \frac{1}{2}\|S_\lambda(A^\top y)\|^2. \tag{12}$$

Using the subgradient inversion formula $\nabla f^* = (\partial f)^{-1}$ [19] we get $\nabla f(y) = S_\lambda(y)$ in view of

$$\partial_i f(x) = \begin{cases} \lambda\,\mathrm{sign}(x_i) + x_i, & x_i \neq 0, \\ [-\lambda, \lambda], & x_i = 0. \end{cases} \tag{13}$$

Thus ψ is unconstrained and differentiable with

$$\nabla\psi(y) = -b + AS_\lambda(A^\top y). \tag{14}$$

Connecting Primal and Dual Variables. In case of a zero duality gap the solutions of (11) and (12) are connected through

$$\overline{x} = S_\lambda(A^\top \overline{y}). \tag{15}$$

We elaborate on this now. With $\mathrm{dom}\, g = 0$, $\mathrm{dom}\, g^* = \mathbb{R}^n$, $\mathrm{dom}\, f^* = \mathbb{R}^n$ and $\mathrm{dom}\, f = \mathbb{R}^n$, the assumptions (50) become $b \in \mathrm{int}\, A(\mathbb{R}^n) = A(\mathrm{int}\, \mathbb{R}^n) =$

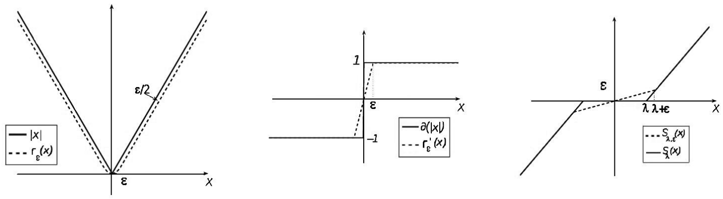

Figure 1: Envelope of $|\cdot|$ (left). Subgradient and gradient of $|\cdot|$ and its envelope (middle). Soft shrinkage operator S_λ and its approximation $S_{\lambda,\varepsilon}$.

$A(\mathbb{R}^n) = \mathcal{R}(A)$, compare [19, Prop. 2.44], and $0 = c \in \operatorname{int}\mathbb{R}^n = \mathbb{R}^n$. Thus, under the assumption $b \in \mathcal{R}(A)$, we have no duality gap. Moreover both problems (11) and (12) have a solution.

Theorem 3.1. *Denote by x_λ and y_λ a solution of* (11) *and* (12) *respectively. Then the following statements are equivalent:*

(a) $b \in \mathcal{R}(A)$, thus the feasible set is nonempty.

(b) The duality gap is zero $\psi(y_\lambda) = \varphi(x_\lambda)$.

(c) Solutions x_λ and y_λ of (11) *and* (12) *exist and are connected through*

$$x_\lambda = S_\lambda(A^\top y_\lambda). \tag{16}$$

Proof. (a) ⇒ (b): holds due to Thm. A.1. On the other hand, (b) implies solvability of ψ and thus (a), in view of the necessary condition $0 = \nabla\psi(y_\lambda) = -b + AS_\lambda(A^\top y_\lambda)$. (a) ⇒ (c): The assumptions of Thm. A.1 hold. Now $\partial f^*(y) = \{\nabla f^*(y)\} = \{S_\lambda(y)\}$ and the r.h.s. of (52a) gives (c). Now, (c) implies $Ax_\lambda = b$ and thus (a). □

The following result shows that for $\lambda \to \infty$ and under the consistency assumption, x_λ given by (16) approaches the ℓ_1-solution of $Ax = b$ (1), if y_λ is a solution of (12). The proof follows along the lines of [21, Prop. 1].

Theorem 3.2. *Denote the solution set of* (1) *by X^*. Assume $X^* \neq \emptyset$ and $b \in \mathcal{R}(A)$. Then for any sequence of positive scalars (λ_k) tending to ∞ and any sequence of vectors (x_{λ_k}), converging to some x^*, we have $x^* \in \operatorname{argmin}_{x\in X^*}\|x\|^2$. If X^* is a singleton, denoted by $\hat{x}$, then $x_{\lambda_k} \to \hat{x}$.*

3.2 Smoothing and Regularization

Primal Problem In order to obtain a strongly convex unconstrained dual problem [19, 1], we regularize the objective by replacing the sparse penalty term through its Moreau envelope so that we obtain a convex differentiable primal cost function with Lipschitz-continuous gradient. Setting

$$r(x) := \|x\|_1, \tag{17}$$

the Moreau envelope reads (cf. (46a))

$$r_\varepsilon(x) := e_\varepsilon r(x) = \sum_{i\in[n]} \begin{cases} \operatorname{sign}(x_i)x_i - \frac{\varepsilon}{2}, & |x_i| > \varepsilon \\ \frac{1}{2\varepsilon}x_i^2, & x_i \in [-\varepsilon, \varepsilon] \end{cases}, \qquad \varepsilon > 0, \tag{18}$$

see Fig. 3.2. We consider the regularised problem

$$\min_{x\in\mathbb{R}^n} f_\varepsilon(x), \quad \text{s.t.} \quad Ax = b, \qquad\qquad f_\varepsilon(x) := \lambda r_\varepsilon(x) + \frac{1}{2}\|x\|^2, \tag{19}$$

which is strongly convex with convexity parameter 1 and differentiable with $\frac{\lambda+\varepsilon}{\varepsilon}$-Lipschitz continuous gradient

$$f_\varepsilon \in \mathcal{S}^{1,1}_{1,\frac{\lambda+\varepsilon}{\varepsilon}}(\mathbb{R}^n), \qquad \partial_i f_\varepsilon(x) = \begin{cases} \lambda \operatorname{sign}(x_i) + x_i, & |x_i| > \varepsilon \\ \frac{\lambda+\varepsilon}{\varepsilon} x_i, & x_i \in [-\varepsilon, \varepsilon] \end{cases}. \tag{20}$$

Dual Problem Writing (19) as

$$\min_{x\in\mathbb{R}^n} f_\varepsilon(x) + \delta_{\{0\}}(b - Ax) \tag{21}$$

we obtain using (49b) the dual problem

$$\min_{y\in\mathbb{R}^m} \psi_\varepsilon(y), \quad \psi_\varepsilon(y) = -\langle b, y\rangle + f^*_\varepsilon(A^\top y). \tag{22}$$

An elementary computation yields

$$f^*_\varepsilon(z) = \sum_{i\in[n]} (f_\varepsilon)^*_i(z_i), \qquad (f_\varepsilon)^*_i(z_i) = \begin{cases} \frac{1}{2}\big(|z_i| - \lambda\big)^2 + \frac{\lambda\varepsilon}{2}, & |z_i| > \lambda + \varepsilon, \\ \frac{1}{2}\frac{\varepsilon}{\lambda+\varepsilon} z_i^2, & |z_i| \le \lambda + \varepsilon \end{cases} \tag{23}$$

and

$$\partial_i f^*_\varepsilon(z) = \begin{cases} \operatorname{sign}(z_i)\left(|z_i| - \lambda\right), & |z_i| > \lambda + \varepsilon, \\ \frac{\varepsilon}{\lambda+\varepsilon} z_i, & |z_i| \le \lambda + \varepsilon. \end{cases} \tag{24}$$

We have $(f_\varepsilon)^* \in \mathcal{S}^{1,1}_{\frac{\varepsilon}{\lambda+\varepsilon},1}(\mathbb{R}^n)$. We denote by $(S_{\lambda,\varepsilon})_i = \partial_i f_\varepsilon^*$ the approximation to the soft thresholding operator, which is numerically S_λ, when $\varepsilon > 0$ is sufficiently small, see Fig. 3.2 left. Thus ψ_ε is unconstrained and differentiable with

$$\nabla\psi_\varepsilon(y) = -b + AS_{\lambda,\varepsilon}(A^\top y), \tag{25}$$

and strongly convex with parameter $\frac{\varepsilon}{\lambda+\varepsilon}$. The solutions of (21) and (22) are connected through

$$\overline{x} = S_{\lambda,\varepsilon}(A^\top \overline{y}). \tag{26}$$

This can be shown in an analogous way as in the previous section.

4 Block Coordinate Descent Type Methods

Consider the problem

$$\min_y \psi(y), \quad \psi \in \mathcal{F}^{1,1}_L(\mathbb{R}^m), \tag{27}$$

and assume that the solution of (27) is nonempty, denoted by Y^*, with corresponding optimal value ψ^*.

We will assume that y has the following partition

$$y = (y_{(1)}, y_{(2)}, \dots, y_{(c)}),$$

where $y_{(i)} \in \mathbb{R}^{m_i}$ and $\sum_{i\in[c]} m_i = m$, $m_i \in \mathbb{N}$. Using the notation from [14], with matrices $U_{(i)} \in \mathbb{R}^{m\times m_i}$, $i \in [c]$ partitioning the identity

$$I = \left(U_{(1)}, U_{(2)}, \dots, U_{(c)}\right),$$

we get

$$y_{(i)} = U_{(i)}^\top y, \quad \forall y \in \mathbb{R}^m,\ \forall i \in [c], \qquad \text{and} \qquad y = \sum_{i\in[c]} U_{(i)} y_{(i)}.$$

Similarly, the vector of partial derivatives corresponding to the variables in $y_{(i)}$ is

$$\nabla_{(i)}\psi(y) = U_{(i)}^\top \nabla\psi(y).$$

We assume the following.

Assumption 4.1. The partial derivatives of ψ are L_i-Lipschitz continuous functions with respect to the block coordinates, that is

$$\|\nabla_{(i)}\psi(y + U_{(i)}h_{(i)}) - \nabla_{(i)}\psi(y)\| \le L_i\|h_{(i)}\|, \quad \forall y,\ h_{(i)} \in \mathbb{R}^{m_i},\ i \in [c]. \tag{28}$$

Lemma 4.1 (Block Descent Lemma)**.** *Suppose that ψ is a continuously differentiable function over $\mathbb{R}^m$ satisfying* (28)*. Then for all $h_{(i)} \in \mathbb{R}^{m_i}$, $i \in [c]$ and $y \in \mathbb{R}^n$ we have*

$$\psi(y + U_{(i)}h_{(i)}) \leq \psi(y) + \langle \nabla_{(i)}\psi(y), h_{(i)}\rangle + \frac{L_i}{2}\|h_{(i)}\|^2. \tag{29}$$

We denote (similarly to [14])

$$R_\alpha(y^{(0)}) = \max_{y \in \mathbb{R}^n}\{\max_{y^* \in Y^*} \|y - y^*\|_\alpha \colon \psi(y) \leq \psi(y^{(0)})\}, \tag{30}$$

where $\alpha = [0, 1]$ and

$$\|y\|_\alpha = \left(\sum_{i \in [c]} L_i^\alpha \|y_{(i)}\|^2\right)^{\frac{1}{2}}. \tag{31}$$

4.1 Randomized Block Coordinate Gradient Descent

The *block coordinate gradient descent* (BCGD) [3] method solves in each iteration the over-approximation in (29)

$$d^{(k,i)} := \text{argmin}_{h_{(i)} \in \mathbb{R}^{n_i}} \langle \nabla_{(i)}\psi(y^{(k)}), h_{(i)}\rangle + \frac{L_i}{2}\|h_{(i)}\|^2, \tag{32}$$

which gives

$$d^{(k,i)} := -\frac{1}{L_i}\nabla_{(i)}\psi(y^{(k)}),$$

and an update of the form

$$y^{(k,i)} := y^{(k)} + d^{(k,i)}. \tag{33}$$

In the deterministic case the next iterate is usually defined after a full cycle through the c blocks. For this particular choice (non-asymptotic) convergence rates were only recently derived in [2], although the convergence of the method was extensively studied in the literature under various assumptions [13, 3]. Instead of using a deterministic cyclic order, randomized strategies were proposed in [14, 12, 16] for choosing a block to update at each iteration of the BCGD method. At iteration k, an index i_k is generated randomly according to the probability distribution vector $p \in \Delta_c$. In [14] the distribution vector was chosen as

$$p_i = \frac{L_i^\alpha}{\sum_{j=1}^c L_j^\alpha}, \quad i \in [c],\ \alpha \in [0, 1]. \tag{34}$$

An expected convergence rate for Alg. 2 was obtained in [14]. We summarize results in the next theorem.

Algorithm 2: Random Block Coordinate Gradient Descent (RBCGD)

```
Input: Starting vector $y^{(0)}$, partition of $[m]$, $U_{(i)}$, $i \in [c]$,
       Lipschitz-constants $L_i$, $p \in \Delta_c$
for $k = 1, 2, \dots$ do
    Sample $i_k \in [c]$ due $i_k \sim p$
    Update $y^{(k+1)} = y^{(k)} - \frac{1}{L_{i_k}} U_{(i_k)} \nabla_{(i_k)} \psi(y^{(k)})$
```

Theorem 4.2. *(Sublinear convergence rate of RBCGD) Let $(y^{(k)})$ be the sequence generated by Alg. 2 and p defined as in* (34)*. Then the expected convergence rate is*

$$\mathbb{E}[\psi(y^{(k)})] - \psi^* \leq \frac{2}{k+4} \left(\sum_{i \in [c]} L_i^{\alpha} \right) R_{1-\alpha}^2(y^{(0)}), \quad k = 0, 1, \dots. \tag{35}$$

In particular, this gives

- *Uniform probabilities* for $\alpha = 0$

$$\mathbb{E}[\psi(y^{(k)})] - \psi^* \leq \frac{2c}{k+4} R_1^2(y^{(0)}), \quad k = 0, 1, \dots$$

- *Probabilities proportional to L_i* for $\alpha = 1$, compare also [17, 18, 20]

$$\mathbb{E}[\psi(y^{(k)})] - \psi^* \leq \frac{2c}{k+4} \left(\frac{1}{c} \sum_{i \in [c]} L_i \right) R_0^2(y^{(0)}), \quad k = 0, 1, \dots$$

As expected, the performance of RBCGD, Alg. 2, improves on strongly convex functions.

Theorem 4.3. *(Linear convergence rate of RBCGD) Let function ψ be strongly convex with respect to the norm $\| \cdot \|_{1-\alpha}$, see* (31)*, with modulus $\mu_{1-\alpha} > 0$. Then, for the $(y^{(k)})$ be the sequence generated by Alg. 2 and the probability vector p from* (34)*, we have*

$$\mathbb{E}[\psi(y^{(k)})] - \psi^* \leq \left(1 - \frac{\mu_{1-\alpha}}{\sum_{i \in [c]} L_i^{\alpha}} \right)^k (\psi(y^{(0)}) - \psi^*), \quad k = 0, 1, \dots. \tag{36}$$

In [2] the authors derive for the deterministic cyclic block coordinate gradient descent convergence rates which were not available before. They also compare the multiplicative constants in the convergence results above to the ones obtained for the deterministic cyclic block order. We refer the interested reader to [2, sec. 3.2].

The above stochastic results from [14] for minimizing convex differentiable functions were generalized in [16] for minimizing the sum of a smooth convex function and a block-separable convex function.

5 Block Sparse Kaczmarz as BCGD

Consider the primal problem (2). Without loss of generality, we consider here a partition of $[m]$ in c blocks, i.e. $\sum_{i\in[c]} m_i = m$, $m_i \in \mathbb{N}$ and $U_{(i)} \in \mathbb{R}^{m\times m_i}$, $i \in [c]$. We define $A_{S_i} = U_{(i)}^\top A$ and denote $A_{(i)} := A_{S_i}$ for simplicity. In the case of non partition $[m] = \cup_{i\in[c]} S_i$, we would consider the extended system (4) along with a partition of $\hat{m}$, see also the beginning of Section 2.

Now recall the iteration of the Randomized Block Sparse Kaczmarz Alg. 1 with stepsize

$$t_k = \frac{1}{\|A_{(i_k)}\|^2},$$

where at iteration k, the block (i_k) is choosen randomly according to the probability distribution vector $p \in \Delta_c$,

$$p_i = \frac{\|A_{(i)}\|^{2\alpha}}{\sum_{j=1}^{c} \|A_{(j)}\|^{2\alpha}}, \quad i \in [c],\ \alpha \in [0,1]. \tag{37}$$

Further consider the randomized block coordinate gradient descent Alg. 2 applied to the dual problem (12).

Proposition 5.1. *The RBSK iteration Alg. 1 where the block $A_{(i_k)}$ is chosen according to $p \in \Delta_c$ from* (37) *is equivalent to the randomized block coordinate gradient descent Alg. 2 applied to the dual function ψ from* (12) *with a stepsize*

$$t_k = \frac{1}{\|A_{(i_k)}\|^2},$$

where i_k is chosen according to $p \in \Delta_c$ from (34), *when for both starting vectors $x^0 = A^\top y^0$ holds.*

Proof. Consider the dual problem (12). The block gradient update applied to ψ from (12) reads

$$y^{(k+1)} = y^{(k)} - \frac{1}{L_{\psi,i}} U_{(i)} \left(A_{(i)} S_\lambda(A^\top y^{(k)}) - b_{(i)} \right), \tag{38}$$

where $L_{\psi,i}$ are the block Lipschitz constants of ψ from (12). These we can compute. Indeed, we have

$$\begin{aligned}\|\nabla_{(i)}\psi(y+U_{(i)}h_{(i)})-\nabla_{(i)}\psi(y)\| &= \|A_{(i)}S_\lambda(A^\top y+A_{(i)}^\top h_{(i)})-A_{(i)}S_\lambda(A^\top y)\| \\ &\leq \|A_{(i)}\|\|S_\lambda(A^\top y+A_{(i)}^\top h_{(i)})-S_\lambda(A^\top y)\| \\ &\leq \|A_{(i)}\|\|A_{(i)}^\top h_{(i)}\| \leq \underbrace{\|A_{(i)}\|^2}_{=:L_{\psi,i}}\|h_{(i)}\|,\end{aligned}$$

due to S_λ being 1-Lipschitz continuous in view of $S_\lambda(x)=\operatorname{prox}_{\lambda\|\cdot\|_1}(x)$.

For some starting point $y^{(0)}$, (38) can be written as

$$z^{(k)} = A^\top y^{(k)}, \tag{39}$$

$$x^{(k)} = S_\lambda(z^{(k)}), \tag{40}$$

$$y^{(k+1)} = y^{(k)} - \frac{1}{L_{\psi,i_k}}U_{(i_k)}\left(A_{(i_k)}x^{(k)}-b_{(i_k)}\right). \tag{41}$$

Thus

$$z^{(k+1)} = A^\top y^{(k+1)} = A^\top y^{(k)} - \frac{1}{L_{\psi,i_k}}A_{(i_k)}^\top\left(A_{(i_k)}x^{(k)}-b_{(i_k)}\right) \tag{42}$$

$$= z^{(k)} - \frac{1}{L_{\psi,i_k}}A_{(i_k)}^\top\left(A_{(i_k)}x^{(k)}-b_{(i_k)}\right). \tag{43}$$

We note that the set S_{i_k} of rows from A defines the indices (i_k) of dual variables y and vice versa. Based on this derivation, the result follows by using mathematical induction. □

Since the sequence generated by the RBCGD method is a sequence of random variables and the efficiency estimate result from Thm. 4.2 bounds the difference of *the expectation* of the function values $\psi(y^{(k)})$ and ψ^* we obtain as a byproduct convergence in expectation.

Theorem 5.2. *Suppose $b \in \mathcal{R}(A)$ holds. Then the randomized RBSK Alg. 1 converges in expectation to the unique solution of* (2).

Proof. In view of Thm. 4.2 the sequence of random variables $\left(\psi(y^{(k)})\right)$ is converging almost surely to ψ^*. The result now follows from Prop. 5.1 and Thm. 3.1. □

Regularized Block Sparse Kaczmarz Method

The above derivation can be repeated for ψ_ε and the corresponding primal problem (19). Consider the dual problem (22) and a partition of $[m]$, $U_{(i)} \in \mathbb{R}^{m \times m_i}$, $i \in [c]$. The block gradient update applied to ψ_ε from (22) reads,

$$y^{(k+1)} = y^{(k)} - \frac{1}{L_{\psi_\varepsilon,i}} U_{(i)} \left(A_{(i)} S_{\lambda,\varepsilon}(A^\top y^{(k)}) - b_{(i)} \right),$$

where $A_{(i)} := U_{(i)}^\top A =: A_{S_i}$ and $L_{\psi_\varepsilon,i}$ are the block Lipschitz constants of ψ_ε from (22). These we can compute in an analogous manner to the Lipschitz constants for ψ and obtain $L_{\psi_\varepsilon,i} = \|A_{(i)}\|^2$, since $S_{\lambda,\varepsilon}$ is as well 1-Lipschitz continuous, see (24).

This leads to the following version of the Block Sparse Kaczmarz Alg. 3. The counterpart of Prop. 5.1 and Thm. 5.2 corresponding to Alg. 3 and RBCGD applied to ψ_ε also hold.

```
Algorithm 3: Regularised RBSK (regRBSK)
Input: Starting vectors $x^{(0)}$ and $z^{(0)}$, covering of $[m] = \cup_{i=1}^{c} S_i$, with
       $|S_i| = m_i$, $i \in [c]$, probability vector $p \in \Delta_c$, choose $\varepsilon > 0$
for $k = 1, 2, \dots$ do
    Sample $i_k \in [c]$ due $i_k \sim p$
    Update $z^{(k+1)} = z^{(k)} - t_k (A_{S_{i_k}})^\top (A_{S_{i_k}} x^{(k)} - b_{S_{i_k}})$
    Update $x^{(k+1)} = S_{\lambda,\varepsilon}(z^{(k)})$
```

Alg. 3 converges faster due to the linear rate in view of the strong convexity. However, for small values of $\varepsilon > 0$ close to zero the estimate in (36) becomes better than the r.h.s. of (35) in Thm. 4.2 only for very high values of k. The strong convexity property of ψ_ε allows to obtain convergence rates in terms of the variables $y^{(k)}$ and $x^{(k)}$. Unfortunately, the tiny value of the convexity parameter will lead to poor convergence rates estimates for these variables too.

Despite these observations, the strong convexity of the dual function ψ_ε opens the way for various accelerations along the lines of [14, 16, 18, 17]. We omit this here, however.

Expected Descent Due to the choice of the distribution (34) we can estimate the expected progress in terms of the expected descent of Alg. 2 in view of both Algorithms 1 and 3 based on primal updates only. Indeed,

$$\begin{aligned}
&\psi(y^{(k)}) - \mathbb{E}[\psi(y^{(k+1)})] \\
&\qquad = \sum_{i_k \in [c]} p_{i_k} \left(\psi(y^{(k)}) - \psi(y^{(k} - \frac{1}{L_{i_k}} U_{(i_k)} \nabla_{(i_k)} \psi(y^{(k)})) \right) \\
&\qquad \geq \sum_{i_k \in [c]} \frac{p_{i_k}}{2L_{\psi,i_k}} \|\nabla_{(i_k)} \psi(y^{(k)})\|^2 \qquad (44) \\
&\qquad = \|\nabla \psi(y^{(k)})\|_w^2 = \|A S_\lambda(A^\top y^{(k)}) - b\|_w^2 \\
&\qquad = \|Ax^{(k)} - b\|_w^2 \qquad (45)
\end{aligned}$$

holds, where $w \in \mathbb{R}^{cm_i}$ is a positive vector and in view of (34) and for every $j \in (i)$

$$w_j = \frac{L_i^{\alpha-1}}{\sum_{j=1}^c L_j^\alpha}, \quad i \in [c], \ \alpha \in [0,1].$$

We now justify the first inequality (44) in the above reasoning. Recall that

$$h_{i_k} = -\frac{1}{L_{\psi,i_k}} \nabla_{(i_k)} \psi(y^{(k)}).$$

minimizes the r.h.s. of (29) for $y = y^{(k)}$. Thus

$$\psi(y^{(k)} + U_{(i_k)} h_{i_k}) \leq \psi(y^{(k)}) - \frac{1}{2L_{\psi,i_k}} \|\nabla_{(i_k)} \psi(y^{(k)})\|^2,$$

and

$$\psi(y^{(k)}) - \psi(y^{(k+1)}) \geq \frac{1}{2L_{\psi,i_k}} \|\nabla_{(i_k)} \psi(y^{(k)})\|^2.$$

This shows (44) and in particular the monotonicity of $\big(\psi(y^{(k)})\big)$. The same computation can be done for ψ_ε.

6 Numerical Experiments

We consider a tomographic reconstruction problem. The goal is to recover an image from line integrals, see Fig. 2 (left). In discretized form, the projection matrix A encodes the incidence geometry of lines and pixels. Each row of A corresponds to a discretized line integral, and each column to a pixel.* Here we consider a binary image, see Fig. 2 (right). However, we don't impose binary constraints here. We only use the prior knowledge that our image is *sparse*. See [8] for a recent textbook on the theory of Compressive Sensing.

*To build the projection matrix A we used the AIRtools package v1.0 [10] obtained from `http://www2.compute.dtu.dk/~pcha/AIRtools/`.

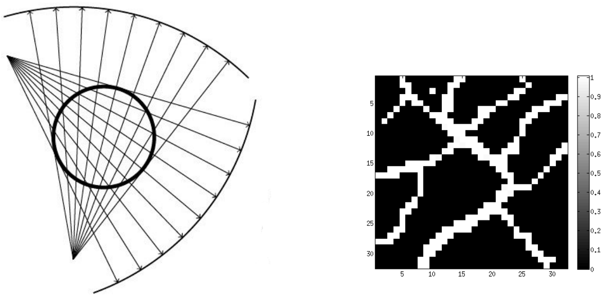

Figure 2: A 32×32 sparse test image (right) is measured from 18 fan beam projections (left) at equiangular source positions.

We refer the interested reader to [15] for further information about when a sufficiently sparse solution can be recovered exactly by (1) in the context of tomography.

6.1 Comparison to Derived Convergence Rates

We first conduct an experiment to illustrate the estimated convergence rates to the empirical rates for different number of blocks from A. We consider the partition of A into c blocks, each with m/c rows if c divides m, or with $\lfloor m/c \rfloor$ except for the first $m \mod (c)$ which will contain $\lfloor m/c \rfloor + 1$ rows. $A_{(i)}$ is the submatrix of A comprising the rows corresponding to the i-th block. In order to find the solution of the perturbed dual formulations we used a conventional unconstrained optimization approach, the *Limited Memory BFGS algorithm* [4], which yields accurate solution approximations and scales to large problem sizes. In all experiments, the perturbation parameters were kept fixed to $\lambda = 10$. We allowed a maximum number of $1000m$ iterations and stopped when the weighted residual $\|Ax^{(k)} - b\|_w^2 \leq 10^{-8}$. The results are summarized in Fig. 3. As expected, a lower number of blocks used leads to fewer iterations. However, more blocks mean cheaper iteration meaning fewer updates for the dual variable or fewer rows used per iteration. Taking this into account we conclude that the fully sequential is the fastest option for the considered example.

6.2 RBSK versus RegRBSK

Having in mind that for ψ_ε the strong convexity parameter equals $\frac{\varepsilon}{\lambda+\varepsilon}$ the convergence is faster the higher the ε is according to (36). However, ε should be small in order that $S_{\lambda,\varepsilon}$ really approximates the soft shrinkage operator S_λ,

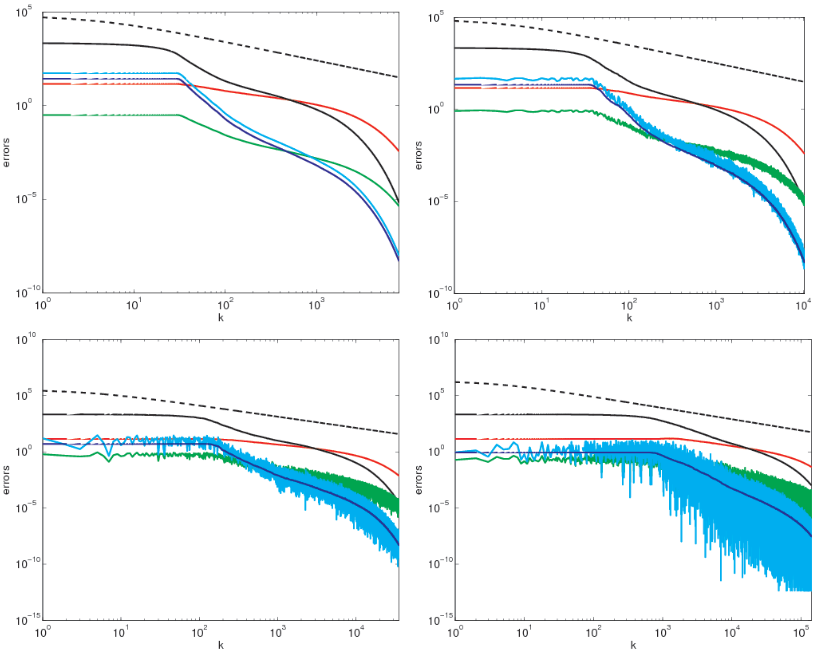

Figure 3: Convergence rates for RBSK and comparison between different errors for different partitions. Here $\alpha = 1$. The empirical error $\mathbb{E}[\psi(y^{(k)})] - \psi^*$ (black curve) is upper bounded by the r.h.s. of (35) (black curve) as estimated by 4.2. We note that both errors $\mathbb{E}[\|y^{(k)} - y^*\|]$ (green curve) and $\mathbb{E}[\|x^{(k)} - x^*\|]$ (red curve) are bounded by the r.h.s. of (35) as well. Interestingly, the difference $\psi(y^{(k+1)}) - \mathbb{E}[\psi(y^{(k+1)})]$ (cyan area) is well approximated by $\|Ax^{(k)} - b\|_w^2$ according to (45). The left upper plot considers the entire matrix A, thus $c = 1$. The right upper plot considers 10 blocks and $c = 10$. The lower plots correspond to $c = 100$ (left) and $c = 710$ (right).

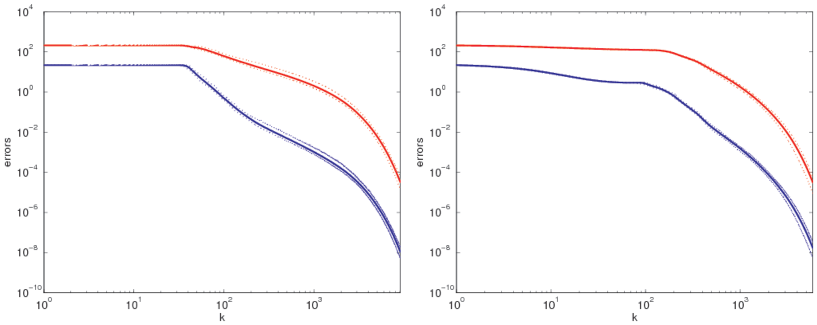

Figure 4: Comparison between RBSK, Alg. 1 and RegRBSK, Alg. 3. RBSK converges (on average) and reaches the tolerance level of 10^{-8} in 9126 iterations, while RegRBSK needs only 5865 iterations. The error $\mathbb{E}[\|x^{(k)} - x^*\|]$ (red curve) and the weighted residual $\mathbb{E}[\|Ax^{(k)} - b\|_w^2]$ (blue curve) is illustrated along with its variance over 100 runs. The number c of partitions was 10.

see Fig. 3.2. We decided for a compromise: slowly decrease $\varepsilon \to 0$ and choose ε iteration dependent, by setting $\varepsilon = \varepsilon_k = (0.99)^k$. The results are illustrated in Fig. 4. This simple technique leads to a significant acceleration.

7 Conclusion and Further Work

We introduced the Randomized Block Sparse Kaczmarz method and the Regularised Randomized Block Sparse Kaczmarz method and showed their expected convergence by exploiting the intimate connection between Bergman projection methods on subsets of linear equations and the coordinate descent method applied on related unconstrained dual problems. The unconstrained duals are obtained by quadratic perturbation of the ℓ_1-objective and of its Moreau envelope. This connection enables to apply existing convergence analysis of the randomized coordinate gradient descent to the (Regularised) Randomized Block Sparse Kaczmarz method. Convergence rates in terms of primal iterates only, are not derived. Experimental results show however that such rates can be observed. Such derivations are the subject of further research.

A Mathematical Background

We collect few definitions and basic facts from [19].

A.1 Proximal Mapping, Moreau Envelope

For a proper, lower-semicontinuous (lsc) function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and parameter value $\lambda > 0$, the *Moreau envelope function* $e_\lambda f$ and the *proximal mapping* $\operatorname{prox}_{\lambda f}(x)$ are defined by

$$e_\lambda f(x) := \inf_y \; f(y) + \frac{1}{2\lambda}\|y - x\|^2, \tag{46a}$$

$$\operatorname{prox}_{\lambda f}(x) := \operatorname*{argmin}_y \; f(y) + \frac{1}{2\lambda}\|y - x\|^2. \tag{46b}$$

We define by

$\mathcal{F}(\mathbb{R}^n)$ the class of convex, proper, lsc functions $f\colon \mathbb{R}^n \to \mathbb{R}$,
$\mathcal{F}^1(\mathbb{R}^n)$ the class of continuous differentiable functions $f \in \mathcal{F}(\mathbb{R}^n)$,
$\mathcal{F}^{1,1}_L(\mathbb{R}^n)$ the class of functions $f \in \mathcal{F}^1(\mathbb{R}^n)$ with Lipschitz-continuous gradient,
$\mathcal{S}^1_\mu(\mathbb{R}^n)$ the class of functions $f \in \mathcal{F}^1(\mathbb{R}^n)$ that are strongly convex with convexity parameter $\mu > 0$,
$\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ the class of functions $f \in \mathcal{F}^{1,1}_L(\mathbb{R}^n) \cap \mathcal{S}^1_\mu(\mathbb{R}^n)$.

For any function $f \in \mathcal{F}$, we have

$$e_\lambda f \in \mathcal{F}^{1,1}_{\frac{1}{\lambda}}, \qquad \nabla e_\lambda f(x) = \frac{1}{\lambda}\big(x - \operatorname{prox}_{\lambda f}(x)\big). \tag{47}$$

Any function $f \in \mathcal{F}$ and its (Legendre-Fenchel) conjugate function $f^* \in \mathcal{F}$ are connected through their Moreau envelopes by

$$(e_\lambda f)^* = f^* + \frac{\lambda}{2}\|\cdot\|^2, \tag{48a}$$

$$\frac{1}{2\lambda}\|x\|^2 = e_\lambda f(x) + e_{\lambda^{-1}} f^*(\lambda^{-1}x), \qquad \forall x \in \mathbb{R}^n, \quad \lambda > 0. \tag{48b}$$

A.2 Fenchel-Type Duality Scheme

Theorem A.1 ([19]). *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $A \in \mathbb{R}^{m \times n}$. Consider the two problems*

$$\inf_{x \in \mathbb{R}^n} \varphi(x), \qquad \varphi(x) = \langle c, x\rangle + f(x) + g(b - Ax), \tag{49a}$$

$$\sup_{y \in \mathbb{R}^m} \psi(y), \qquad \psi(y) = \langle b, y\rangle - g^*(y) - f^*(A^\top y - c) \;\; . \tag{49b}$$

where the functions f and g are proper, lower-semicontinuous (lsc) and convex. Suppose that

$$b \in \operatorname{int}(A \operatorname{dom} f + \operatorname{dom} g), \tag{50a}$$

$$c \in \operatorname{int}(A^\top \operatorname{dom} g^* - \operatorname{dom} f^*) \ . \tag{50b}$$

Then the optimal solutions $\overline{x}, \overline{y}$ are determined by

$$0 \in c + \partial f(\overline{x}) - A^\top \partial g(b - A\overline{x}), \qquad 0 \in b - \partial g^*(\overline{y}) - A \partial f^*(A^\top \overline{y} - c) \tag{51a}$$

and connected through

$$\overline{y} \in \partial g(b - A\overline{x}), \qquad \overline{x} \in \partial f^*(A^\top \overline{y} - c), \tag{52a}$$

$$A^\top \overline{y} - c \in \partial f(\overline{x}), \qquad b - A\overline{x} \in \partial g^*(\overline{y}) \ . \tag{52b}$$

Acknowledgement. SP gratefully acknowledges financial support from the Ministry of Science, Research and Arts, Baden-Württemberg, within the Margarete von Wrangell postdoctoral lecture qualification program.

References

[1] H. H. Bauschke and P. L. Combettes. The Baillon-Haddad Theorem Revisited. *Journal of Convex Analysis*, 17:781–787, 2010.

[2] A. Beck and L. Tetruashvili. On the Convergence of Block Coordinate Descent Type Methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[3] D. P. Bertsekas. *Nonlinear Programming.* Belmont MA: Athena Scientific, 2nd edition, 1999.

[4] J.-F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization – Theoretical and Practical Aspects.* Springer Verlag, Berlin, 2006.

[5] C. Byrne and Y. Censor. Proximity Function Minimization Using Multiple Bregman Projections, with Applications to Split Feasibility and Kullback-Leibler Distance Minimization. *Annals of Operations Research*, 105(1-4):77–98, 2001.

[6] J.-F. Cai, S. Osher, and Z. Shen. Convergence of the Linearized Bregman Iteration for l1-norm Minimization. *Mathematics of Computation*, 78(268):2127–2136, 2009.

[7] Y. Censor and T. Elfving. A multiprojection algorithm using Bregman projections in a product space. *Numerical Algorithms*, 8(2):221–239, 1994.

[8] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[9] M. P. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM Journal on Optimization*, 18(4):1326–1350, 2007.

[10] P.C. Hansen and M. Saxild-Hansen. {AIR} tools a {MATLAB} package of algebraic iterative reconstruction methods. *Journal of Computational and Applied Mathematics*, 236(8):2167 – 2178, 2012. Inverse Problems: Computation and Applications.

[11] D. A. Lorenz, F. Schöpfer, and S. Wenger. The Linearized Bregman Method via Split Feasibility Problems: Analysis and Generalizations. *SIAM Journal on Imaging Science*, 7(2):1237–1262, 2014.

[12] Z. Lu and L. Xiao. On the Complexity Analysis of Randomized Block-Coordinate Descent Methods. *CoRR*, 2013.

[13] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, January 1992.

[14] Y. Nesterov. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[15] S. Petra and C. Schnörr. Average Case Recovery Analysis of Tomographic Compressive Sensing. *Linear Algebra and its Applications*, 441:168–198, 2014. Special issue on Sparse Approximate Solution of Linear Systems.

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

[17] P. Richtárik and M. Takác. Parallel Coordinate Descent Methods for Big Data Optimization. *CoRR*, abs/1212.0873, 2012.

[18] P. Richtárik and M. Takác. On Optimal Probabilities in Stochastic Coordinate Descent Methods. *CoRR*, abs/1310.3438, 2013.

[19] R.T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2nd edition, 2009.

[20] T. Strohmer and R. Vershynin. A Randomized Kaczmarz Algorithm with Exponential Convergence. *Journal of Fourier Analysis and Applications*, 15:262–278, 2009.

[21] P. Tseng. Convergence and Error Bound for Perturbation of Linear Programs. *Computational Optimization and Applications*, 13(1-3):221–230, 1999.

[22] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman Iterative Algorithms for l1-Minimization with Applications to Compressed Sensing. *SIAM Journal on Imaging Sciences*, pages 143–168, 2008.

Stefania Petra,
Image and Pattern Analysis Group,
Department of Mathematics and Computer Science,
University of Heidelberg,
Speyerer Str. 6, 69115 Heidelberg, Germany
Email:petra@math.uni-heidelberg.de