Operations Research Letters 44 (2016) 61–66

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Rate of convergence analysis of dual-based variables decomposition methods for strongly convex problems

Amir Beck*, Luba Tetruashvili, Yakov Vaisbourd, Ariel Shemtov

Faculty of Industrial Engineering and Management, Technion, Haifa, Israel

ARTICLE INFO

Article history:
Received 28 July 2015
Received in revised form
15 November 2015
Accepted 15 November 2015
Available online 2 December 2015

Keywords:
Dual based methods
Block variables decomposition
Total variation denoising

ABSTRACT

We consider the problem of minimizing the sum of a strongly convex function and a term comprising the sum of extended real-valued proper closed convex functions. We derive the primal representation of dual-based block descent methods and establish a relation between primal and dual rates of convergence, allowing to compute the efficiency estimates of different methods. We illustrate the effectiveness of the methods by numerical experiments on total variation-based denoising problems.

1. Introduction

1.1. The basic setting

In this paper, our aim is to devise simple methods for solving minimization problems of the form

$$(\mathrm{P})\ \min_{\mathbf{x}\in\mathbb{E}}\left\{f(\mathbf{x})+\sum_{i=1}^{m}\psi_i(\mathbf{x})\right\},$$

with $\mathbb{E}$ being a given final dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and associated Euclidean norm $\|\mathbf{x}\| \equiv \sqrt{\langle\mathbf{x},\mathbf{x}\rangle}$. The functions f and ψ_i satisfy the following conditions that are summarized in one assumption.

Assumption 1.
- $f:\mathbb{E}\to(-\infty,\infty]$ is a closed, proper extended valued σ-strongly convex function.
- $\psi_i:\mathbb{E}\to(-\infty,\infty]$ $(i=1,2,\ldots,m)$ is a closed, proper extended real-valued convex function.
- $\mathrm{ri}(\mathrm{dom}\, f)\cap\left(\bigcap_{i=1}^{m}\mathrm{ri}(\mathrm{dom}\,\psi_i)\right)\neq\emptyset$.

Under the latter assumption, problem (P) has a unique minimizer that we denote by $\mathbf{x}^*$; the optimal value is denoted by $f_{\mathrm{opt}}=f(\mathbf{x}^*)$. The dual problem of (P) is given by

$$(\mathrm{D})\ \max_{\mathbf{y}}\left\{q(\mathbf{y})\equiv -f^*\left(-\sum_{j=1}^{m}\mathbf{y}_j\right)-\sum_{j=1}^{m}\psi_j^*(\mathbf{y}_j)\right\},\tag{1.1}$$

where $f^*(\cdot)=\sup_{\mathbf{x}\in\mathbb{E}}\langle\cdot,\mathbf{x}\rangle-f(\mathbf{x})$ and $\psi_i^*(\cdot)=\sup_{\mathbf{x}\in\mathbb{E}}\langle\cdot,\mathbf{x}\rangle-\psi_i(\mathbf{x})$ are the corresponding conjugate functions. The duality between (P) and (D) is obviously a simple application of Fenchel's (as well as Lagrangian) duality [18]. In this specific form, it is also known as the duality between the regularized consensus problem and the sharing problem (see Section 7 of [7]).

Since Slater's condition is satisfied, and since the primal problem is bounded below, strong duality holds, which means that the optimal solution of the dual problem is attained and the optimal value of the dual problem, which we denote by q_{opt}, coincides with the primal optimal value:

$$f_{\mathrm{opt}}=\mathrm{val}\,(P)=\mathrm{val}\,(D)=q_{\mathrm{opt}}.$$

Using the notation $\mathbf{y}=(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_m)$, the dual problem (D) in minimization form can be written as

$$\min_{\mathbf{y}\in\mathbb{E}^m}\left\{H(\mathbf{y})\equiv F(\mathbf{y})+\sum_{i=1}^{m}\Psi_i(\mathbf{y}_i)\right\}\tag{1.2}$$

with

$$F(\mathbf{y})\equiv f^*\left(-\sum_{j=1}^{m}\mathbf{y}_j\right),\qquad \Psi_j(\mathbf{y}_j)\equiv\psi_j^*(\mathbf{y}_j).\tag{1.3}$$

Under Assumption 1, $\Psi_1,\Psi_2,\ldots,\Psi_m$ are closed, proper and convex and, by the well-known Baillon–Haddad Lemma (see

* Corresponding author.
E-mail addresses: becka@ie.technion.ac.il (A. Beck), lubate@campus.technion.ac.il (L. Tetruashvili), yakovv@campus.technion.ac.il (Y. Vaisbourd), ariel.shemtov@ensta-paristech.fr (A. Shemtov).

http://dx.doi.org/10.1016/j.orl.2015.11.007
0167-6377/

[19, Section 12H]), F is an L-smooth function with $L = \frac{m}{\sigma}$, meaning that $\|\nabla F(\mathbf{y}) - \nabla F(\mathbf{w})\| \leq L\|\mathbf{y} - \mathbf{w}\|$ for any $\mathbf{y}, \mathbf{w} \in \mathbb{E}$. In addition, for any i, $\nabla_i F$ is Lipschitz continuous with constant $\frac{1}{\sigma}$. The optimal solution set of the dual problem will be denoted by Y^*.

1.2. Paper layout

The main objective of the paper is to present a convergence analysis of dual-based decomposition methods for solving (P), where the basic step in the dual algorithm will either consist of a well known exact minimization [10] or a proximal gradient step [3,9] with respect to the corresponding block of dual variables. We begin in Section 2 by deriving a primal representation of both dual block proximal gradient and dual alternating minimization methods. We then establish in Section 3 a relation between certain primal and dual distances to optimality that allows to automatically translate any rate of convergence result in the dual space into a rate of convergence result in the primal space. We then utilize known results on rates of convergence for variables decomposition methods in order to establish new corresponding results for dual-based decomposition methods. Finally, we demonstrate in Section 4 the potential of the derived methods in the context of total variation-based denoising problems.

1.3. Literature review

Variables decomposition methods such as the alternating minimization method were extensively studied for many years, see e.g., [1,6,16,10]. Rate of convergence results under certain strong convexity and/or error bound assumptions were established in [10,13]. The first rate of convergence result in the deterministic setting without any strong convexity/error bounds assumption was established in [5], where an $O(1/k)$ rate convergence of the block coordinate gradient projection method was shown. In the unconstrained case, it was shown that the method can be accelerated to a rate of $O(1/k^2)$. The work [5] also established an $O(1/k)$ rate of convergence for the alternating minimization method with two blocks in the smooth unconstrained case with a multiplicative constant that depends on the minimum of the block Lipschitz constants. The latter was later generalized in [2] to the case of a composite objective function with a separable nonsmooth paper. Recently, it was shown in [11] that a sublinear rate of convergence can also be established for the block proximal gradient and alternating minimization methods with arbitrary number of blocks. Randomized methods in which the blocks are not picked by a deterministic rule, but rather by some random distribution on the indices set are also the topic of an extensive research [15,17,12].

The idea of solving a problem of the form (P) via a dual-based block decomposition method for the case $m = 2$ was studied in [8].

2. Dual-based block descent methods

2.1. Step types

We begin by describing the two types of minimization operations that will be employed on a given block $i \in \{1, 2, \ldots, m\}$. We assume that the dual variables are given by $\mathbf{y}_j = \bar{\mathbf{y}}_j$, $j \in \{1, 2, \ldots, m\}$, and show how to compute the new value of $\mathbf{y}_i$, which we denote by $\mathbf{y}_i^{\text{new}}$. We consider two options for the dual step employed on the ith block:

- **dual exact minimization step**.

$$\mathbf{y}_i^{\text{new}} \in \operatorname*{argmin}_{\mathbf{y}_i} \left\{ f^*\left(-\sum_{j=1, j\neq i}^{m} \bar{\mathbf{y}}_j - \mathbf{y}_i \right) + \psi_i^*(\mathbf{y}_i) \right\}. \tag{2.1}$$

  Note that for this minimization step, the value of $\bar{\mathbf{y}}_i$ is not being used.
- **dual proximal gradient step**.

$$\mathbf{y}_i^{\text{new}} = \operatorname{prox}_{\sigma\psi_i^*}\left(\bar{\mathbf{y}}_i + \sigma \nabla f^*\left(-\sum_{j=1}^{m} \bar{\mathbf{y}}_j \right) \right). \tag{2.2}$$

2.1.1. Primal representation of the dual exact minimization step

To derive a primal representation of (2.1), let us write it as

$$\min_{\mathbf{y}_i, \mathbf{w}} \left\{ f^*(\mathbf{w}) + \psi_i^*(\mathbf{y}_i) : \mathbf{w} + \mathbf{y}_i = -\tilde{\mathbf{y}}_i \right\}, \tag{2.3}$$

where $\tilde{\mathbf{y}}_i = \sum_{j=1,\ j\neq i}^{m} \bar{\mathbf{y}}_j$. The dual problem of (2.3) is

$$\begin{aligned}
&\max_{\mathbf{x}} \min_{\mathbf{w}, \mathbf{y}_i} \left\{ f^*(\mathbf{w}) + \psi_i^*(\mathbf{y}_i) - \langle \mathbf{x}, \mathbf{w} + \mathbf{y}_i + \tilde{\mathbf{y}}_i \rangle \right\} \\
&= \max_{\mathbf{x}} \left\{ \left[\min_{\mathbf{w}} (f^*(\mathbf{w}) - \langle \mathbf{x}, \mathbf{w} \rangle) \right] \right. \\
&\qquad \left. + \left[\min_{\mathbf{y}_i} (\psi_i^*(\mathbf{y}_i) - \langle \mathbf{x}, \mathbf{y}_i \rangle) \right] - \langle \mathbf{x}, \tilde{\mathbf{y}}_i \rangle \right\} \\
&= \max_{\mathbf{x}} \left\{ -f(\mathbf{x}) - \psi_i(\mathbf{x}) - \langle \mathbf{x}, \tilde{\mathbf{y}}_i \rangle \right\},
\end{aligned}$$

where in the last equality we used the fact that $f = f^{**}$ and $\psi_i = \psi_i^{**}$ (since f and ψ_i are closed, proper and convex). We can thus conclude that $\mathbf{y}_i^{\text{new}}$ can be determined by first computing $\bar{\mathbf{x}} \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \psi_i(\mathbf{x}) + \langle \tilde{\mathbf{y}}_i, \mathbf{x} \rangle \right\}$, and then choosing $\mathbf{y}_i^{\text{new}} \in \operatorname{argmax}_{\mathbf{y}_i} \left\{ \langle \mathbf{y}_i, \bar{\mathbf{x}} \rangle - \psi_i^*(\mathbf{y}_i) \right\}$, which is exactly the same as $\mathbf{y}_i^{\text{new}} \in \partial\psi_i(\bar{\mathbf{x}})$. Therefore, step (2.1) is equivalent to

Primal representation of the dual exact minimization step:

$$\bar{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \psi_i(\mathbf{x}) + \langle \tilde{\mathbf{y}}_i, \mathbf{x} \rangle \right\}, \quad \left(\tilde{\mathbf{y}}_i = \sum_{j\neq i} \bar{\mathbf{y}}_j \right) \tag{2.4}$$

$$\mathbf{y}_i^{\text{new}} \in \partial\psi_i(\bar{\mathbf{x}}). \tag{2.5}$$

When f is also assumed to be continuously differentiable over $\mathbb{E}$, we can use the first-order optimality condition on problem (2.4) to conclude that $-\nabla f(\bar{\mathbf{x}}) - \tilde{\mathbf{y}}_i \in \partial\psi_i(\bar{\mathbf{x}})$. Therefore, step (2.5) can be replaced (in this setting) with $\mathbf{y}_i^{\text{new}} = -\nabla f(\bar{\mathbf{x}}) - \tilde{\mathbf{y}}_i$.

2.1.2. Primal representation of the dual proximal gradient step

To find a primal representation of (2.2), first note that

$$\begin{aligned}
\nabla f^*\left(-\sum_{j=1}^{m} \bar{\mathbf{y}}_j \right) &= \operatorname*{argmax}_{\mathbf{x}\in\mathbb{E}} \left\{ \left\langle -\sum_{j=1}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \right\rangle - f(\mathbf{x}) \right\} \\
&= \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \left\langle \sum_{j=1}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \right\rangle \right\}.
\end{aligned}$$

By denoting the above argmin/argmax by

$$\bar{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \left\langle \sum_{j=1}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \right\rangle \right\},$$

we obtain that the proximal gradient step takes the form $\mathbf{y}_i^{\text{new}} = \operatorname{prox}_{\sigma\psi_i^*}(\bar{\mathbf{y}}_i + \sigma\bar{\mathbf{x}})$. Using the Moreau decomposition formula [14], $\operatorname{prox}_{\sigma\psi_i^*}(\mathbf{z}) = \mathbf{z} - \sigma \operatorname{prox}_{\psi_i/\sigma}(\mathbf{z}/\sigma)$, and hence,

$$\mathbf{y}_i^{\text{new}} = \bar{\mathbf{y}}_i + \sigma\bar{\mathbf{x}} - \sigma \operatorname{prox}_{\psi_i/\sigma}\left(\frac{\bar{\mathbf{y}}_i}{\sigma} + \bar{\mathbf{x}} \right).$$

Therefore, step (2.2) can be rewritten as

Primal representation of the dual proximal gradient step:

$$\bar{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \langle \sum_{j=1}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \rangle \right\},$$

$$\bar{\mathbf{y}}_i^{\text{new}} = \bar{\mathbf{y}}_i + \sigma \bar{\mathbf{x}} - \sigma \operatorname{prox}_{\psi_i/\sigma} \left(\frac{\bar{\mathbf{y}}_i}{\sigma} + \bar{\mathbf{x}} \right).$$

2.2. Methods

Using the primal presentations of the dual block descent steps, we can now write explicitly the dual alternating minimization and dual block proximal gradient methods.

Dual Alternating Minimization. We begin by describing the dual alternating minimization method.

Dual Cyclic Alternating Minimization Method (DAM-C)
Initialization. $\mathbf{y}^0 = (\mathbf{y}_0^0, \mathbf{y}_1^0, \ldots, \mathbf{y}_m^0) \in \mathbb{E}^m$.
General Step ($k = 0, 1, 2, 3, \ldots$).

- Set $\mathbf{y}^{k,0} = \mathbf{y}^k$.
- For $i = 0, 1, \ldots, m-1$, define $\mathbf{y}^{k,i+1}$ as follows:

$$\mathbf{x}^{k,i} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \psi_{i+1}(\mathbf{x}) + \langle \textstyle\sum_{j=1, j\neq i+1}^{m} \mathbf{y}_j^{k,i}, \mathbf{x} \rangle \right\}$$

$$\mathbf{y}_j^{k,i+1} \begin{cases} \in \partial \psi_{i+1}(\mathbf{x}^{k,i}) & j = i+1, \\ = \mathbf{y}_j^{k,i} & j \neq i+1. \end{cases}$$

- Set $\mathbf{y}^{k+1} = \mathbf{y}^{k,m}$ and $\mathbf{x}^k = \mathbf{x}^{k,0}$.

As before, if $f \in C^1$, then the update rule for $\mathbf{y}_{i+1}^{k,i+1}$ can be replaced by $\mathbf{y}_{i+1}^{k,i+1} = -\nabla f(\mathbf{x}^{k,i}) - \sum_{j=1, j\neq i+1}^{m} \mathbf{y}_j^{k,i}$. The above scheme uses a cyclic index selection strategy, which is obviously a deterministic rule. Another strategy which is quite common [15] is randomized index selection strategy in which at each iteration an index is picked according to a uniform distribution.

Dual Block Proximal Gradient

In the dual block proximal gradient method, at each iteration a proximal gradient step is performed in the dual space. Below we describe the variant in which the index is chosen in a cyclic manner.

Dual Cyclic Block Proximal Gradient Method (DBPG-C)
Initialization. $(\mathbf{y}_0^0, \mathbf{y}_1^0, \ldots, \mathbf{y}_m^0) \in \mathbb{E}^m$.
General Step ($k = 0, 1, 2, 3, \ldots$).

- Set $\mathbf{y}^{k,0} = \mathbf{y}^k$.
- For $i = 0, 1, \ldots, m-1$, define $\mathbf{y}^{k,i+1}$ as follows

$$\mathbf{x}^{k,i} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \langle \textstyle\sum_{j=1}^{m} \mathbf{y}_j^{k,i}, \mathbf{x} \rangle \right\},$$

$$\mathbf{y}_j^{k,i+1} = \begin{cases} \mathbf{y}_{i+1}^k + \sigma \mathbf{x}^{k,i} \\ \quad -\sigma \operatorname{prox}_{\psi_{i+1}/\sigma} \left(\frac{\mathbf{y}_{i+1}^k}{\sigma} + \mathbf{x}^{k,i} \right) & j = i+1, \\ \mathbf{y}_j^{k,i}, & j \neq i+1. \end{cases}$$

- Set $\mathbf{y}^{k+1} = \mathbf{y}^{k,m}$ and $\mathbf{x}^k = \mathbf{x}^{k,0}$.

Note that in both DAM-C and DBPG-C we use a cyclic index selection rule. In the sequel we will also refer to the randomized versions, DBPG-R, in which at each iteration an index is picked at random via a uniform distribution.

3. Convergence analysis

3.1. The primal–dual rate relation

Suppose that in the dual space we have an arbitrary algorithm that produces points $\{\mathbf{y}^k\}$ satisfying

$$q_{\text{opt}} - q(\mathbf{y}^k) \leq \theta(k),$$

where $\theta(k)$ is the so-called *efficiency estimate*, which goes to 0 as k tends to ∞. We would like to establish a rate of convergence of the primal sequence, which is denoted by $\{\mathbf{x}^k\}$ in each of the described algorithms. Note that the definition of $\mathbf{x}^k$, given the dual variables vector $\mathbf{y}^k$ is not the same for the DBPG and DAM methods. The formula for $\mathbf{x}^k$ in the DBPG method is

$$\mathbf{x}^k = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}) + \left\langle \sum_{j=1}^{m} \mathbf{y}_j^k, \mathbf{x} \right\rangle \right\}, \tag{3.1}$$

while the formula for $\mathbf{x}^k$ in the DAM method is

$$\mathbf{x}^k = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}) + \psi_{i_k}(\mathbf{x}) + \left\langle \sum_{j\neq i_k} \mathbf{y}_j^k, \mathbf{x} \right\rangle \right\} \tag{3.2}$$

for some $i_k \in \{1, 2, \ldots, m\}$. Note that the above holds also for the randomized versions of the algorithms.

Theorem 3.1 (*Primal–Dual Relation*). *Let $\bar{\mathbf{y}}$ satisfy $\bar{\mathbf{y}}_j \in \operatorname{dom} \psi_j^*$ for any $j \in \{1, 2, \ldots, m\}$. Let $\bar{\mathbf{x}}$ be defined by either*

$$\bar{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}) + \left\langle \sum_{i=j}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \right\rangle \right\} \tag{3.3}$$

or

$$\bar{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}) + \psi_i(\mathbf{x}) + \left\langle \sum_{j=1, j\neq i}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \right\rangle \right\} \tag{3.4}$$

for some $i \in \{1, 2, \ldots, m\}$. Then

$$\|\bar{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\text{opt}} - q(\bar{\mathbf{y}})). \tag{3.5}$$

Before proving the theorem, we note that for the case where $\bar{\mathbf{x}}$ is defined by (3.3), the result was already established (although not explicitly stated) as part of the proof of [4, Theorem 4.1]. We thus only prove the result for the case where $\bar{\mathbf{x}}$ is defined by (3.4).

Proof. Assume that $\bar{\mathbf{x}}$ is defined by (3.4). Define the vectors $\bar{\mathbf{z}}_j$ as

$$\bar{\mathbf{z}}_j \in \operatorname*{argmin}_{\mathbf{z}\in\mathbb{E}} \left\{ \psi_j(\mathbf{z}) - \langle \bar{\mathbf{y}}_j, \mathbf{z} \rangle \right\}, \quad j = 1, 2, \ldots, m. \tag{3.6}$$

For any $k \geq 0$, we will define

$$\tilde{h}(\mathbf{x}) \equiv f(\mathbf{x}) + \psi_i(\mathbf{x}) + \left\langle \sum_{j=1, j\neq i}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \right\rangle$$

$$\tilde{s}(\mathbf{z}_{[i]}) = \tilde{s}(\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_m) \equiv \sum_{j=1, j\neq i}^{m} [\psi_j(\mathbf{z}_j) - \langle \bar{\mathbf{y}}_j, \mathbf{z}_j \rangle],$$

where here we use the notation that given a vector $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) \in \mathbb{E}^m$, then for any $k \in \{1, 2, \ldots, m\}$, we define

$\mathbf{v}_{[k]} \in \mathbb{E}^{m-1}$ as $\mathbf{v}_{[k]} = (\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_m)$. Let us consider the following Lagrangian function ($\mathbf{x} \in \operatorname{dom} f$, $\mathbf{z}_j \in \operatorname{dom} \psi_j$):

$$\begin{aligned} L_i(\mathbf{x}, \mathbf{z}; \bar{\mathbf{y}}) &= f(\mathbf{x}) + \psi_i(\mathbf{x}) + \sum_{j=1, j\neq i}^{m} \psi_j(\mathbf{z}_j) + \sum_{j=1, j\neq i}^{m} \langle \bar{\mathbf{y}}_j, \mathbf{x} - \mathbf{z}_j \rangle \\ &= f(\mathbf{x}) + \psi_i(\mathbf{x}) + \left\langle \sum_{j=1, j\neq i}^{m} \bar{\mathbf{y}}_j, \mathbf{x} \right\rangle \\ &\quad + \sum_{j=1, j\neq i}^{m} [\psi_j(\mathbf{z}_j) - \langle \bar{\mathbf{y}}_j, \mathbf{z}_j \rangle]. \end{aligned}$$

Note that the Lagrangian function L_i does not depend on the input vector $\mathbf{y}_i$ and the Lagrange multipliers vector $\bar{\mathbf{y}}_i$. The definition of L_i readily implies that

$$L_i(\mathbf{x}, \mathbf{z}; \bar{\mathbf{y}}) = \tilde{h}(\mathbf{x}) + \tilde{s}(\mathbf{z}_{[i]}).$$

By the σ-strong convexity of $\tilde{h}$ and the definition of $\bar{\mathbf{z}}_j$ and $\bar{\mathbf{x}}$ (see (3.6)), we have $\tilde{h}(\mathbf{x}) - \tilde{h}(\bar{\mathbf{x}}) \geq \frac{\sigma}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|^2$ and $\tilde{s}(\mathbf{z}_{[i]}) - \tilde{s}(\bar{\mathbf{z}}_{[i]}) \geq 0$. Summing the two inequalities, we obtain that for any $\mathbf{x} \in \operatorname{dom} f$, $\mathbf{z} \in \operatorname{dom} \psi_1 \times \cdots \times \operatorname{dom} \psi_m$:

$$L_i(\mathbf{x}, \mathbf{z}; \bar{\mathbf{y}}) - L_i(\bar{\mathbf{x}}, \bar{\mathbf{z}}; \bar{\mathbf{y}}) \geq \frac{\sigma}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|^2. \tag{3.7}$$

Defining $\mathbf{z}_j^* = \mathbf{x}^*$, and using the optimality of $\mathbf{x}^*$, we obtain that

$$L_i(\mathbf{x}^*, \mathbf{z}^*; \bar{\mathbf{y}}) = q_{\text{opt}}. \tag{3.8}$$

In addition,

$$\begin{aligned} q(\bar{\mathbf{y}}) &= \min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + \sum_{j=1}^{m} \left[\psi_j(\mathbf{z}_j) + \langle \bar{\mathbf{y}}_j, \mathbf{x} - \mathbf{z}_j \rangle\right] \\ &\leq \min_{\mathbf{x},\mathbf{z}} \left\{ f(\mathbf{x}) + \sum_{j=1}^{m} \left[\psi_j(\mathbf{z}_j) + \langle \bar{\mathbf{y}}_j, \mathbf{x} - \mathbf{z}_j \rangle\right] : \mathbf{z}_i = \mathbf{x} \right\} \\ &= \min_{\mathbf{x},\mathbf{z}_{[i]}} \left\{ f(\mathbf{x}) + \psi_i(\mathbf{x}) + \sum_{j=1, j\neq i}^{m} \left[\psi_j(\mathbf{z}_j) + \langle \bar{\mathbf{y}}_j, \mathbf{x} - \mathbf{z}_j \rangle\right] \right\} \\ &= L_i(\bar{\mathbf{x}}, \bar{\mathbf{z}}; \bar{\mathbf{y}}). \end{aligned} \tag{3.9}$$

Thus, plugging $\mathbf{x} = \mathbf{x}^*, \mathbf{z} = \mathbf{z}^*$ in (3.7) and using (3.8) and the inequality (3.9), the desired inequality (3.5) follows. □

We can readily write a result on the rate of convergence of the primal sequence, given an efficiency estimate for the dual sequence.

Theorem 3.2. *Let $\{\mathbf{y}^k\}_{k\geq 0}$ be any sequence satisfying $\mathbf{y}_j^k \in \operatorname{dom} \psi_j^*$, $j = 1, 2, \ldots, m$, and assume that $\mathbf{x}^k$ is defined by either (3.1) or by (3.2) for some $i \in \{1, 2, \ldots, m\}$. If*

$$q_{\text{opt}} - q(\mathbf{y}^k) \leq \theta(k),$$

where $\theta : \mathbb{N} \to \mathbb{R}_{++}$ is any positive-valued function of natural numbers. Then

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}\theta(k).$$

When the algorithm is not deterministic, the result is replaced by a result with expectations.

Theorem 3.3. *Let $\{\mathbf{y}^k\}_{k\geq 0}$ be a sequence of random variables satisfying $\mathbf{y}_j^k \in \operatorname{dom} \psi_j^*$ for any $j \in \{1, 2, \ldots, m\}$ and $k \geq 0$, and assume that $\mathbf{x}^k$ is defined by either (3.1) or by (3.2) for some $i_k \in \{1, 2, \ldots, m\}$. Suppose that $q^* - \mathbb{E}(q(\mathbf{y}^k)) \leq \theta(k)$. Then $\mathbb{E}\left[\|\mathbf{x}^k - \mathbf{x}^*\|^2\right] \leq \frac{2}{\sigma}\theta(k)$.*

3.2. Rate of convergence of variables decomposition methods

As was noted in the introduction, the rates of convergence of various variables decomposition methods have been established in the literature. In the table below, known efficiency estimates are summarized in the terminology used in this paper.

Method	Complexity result	Remarks	Source
BPG-C	$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{C_1}{k+1}$	Ψ_i^* indicators, general m	[5]
BPG-C	$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{C_2}{k+1}$	General Ψ_i and m	[11]
BPG-R	$q_{\text{opt}} - \mathbb{E}(q(\mathbf{y}^k)) \leq \frac{mC_3}{m+k}$	General Ψ_i and m	[12]
AM-C	$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{C_4}{k}$	$m = 2$	[2]
AM-C	$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{C_5}{k+1}$	General Ψ_i and m	[11]

The constants that appear in the efficiency estimates are:

$$C_1 = \frac{3m\left[(2m+1)R + \sigma M\right]^2}{\sigma},$$

$$C_2 = 2\sigma m G_{\max}^2 R^2 \max\left\{\frac{2}{\sigma m G_{\max}^2 R^2} - 2, q_{\text{opt}} - q(\mathbf{y}^0), 2\right\},$$

$$C_3 = \frac{1}{2\sigma}\min_{\mathbf{y}^* \in Y^*} \|\mathbf{y}_0 - \mathbf{y}^*\|^2 + q_{\text{opt}} - q(\mathbf{y}^0),$$

$$C_4 = 3\max\left\{q_{\text{opt}} - q(\mathbf{y}^0), \frac{1}{\sigma}R^2\right\},$$

$$C_5 = \frac{2m^2R^2\max\left\{\frac{2\sigma}{m^2R^2} - 2, q_{\text{opt}} - q(\mathbf{y}^0), 2\right\}}{\sigma}$$

with $M = \max_{\mathbf{y}^* \in Y^*} \|\nabla F(\mathbf{y}^*)\|$ and $R = \max_{\mathbf{y} \in \mathbb{E}^m, \mathbf{y}^* \in Y^*}\{\|\mathbf{y} - \mathbf{y}^*\| : q(\mathbf{y}) \geq q(\mathbf{y}^0)\}$. The constant $G_{\max}$ is an upper bound on the maximum of several Lipschitz constants of the gradient of an upper bounding function of q (more details can be found in [11]).

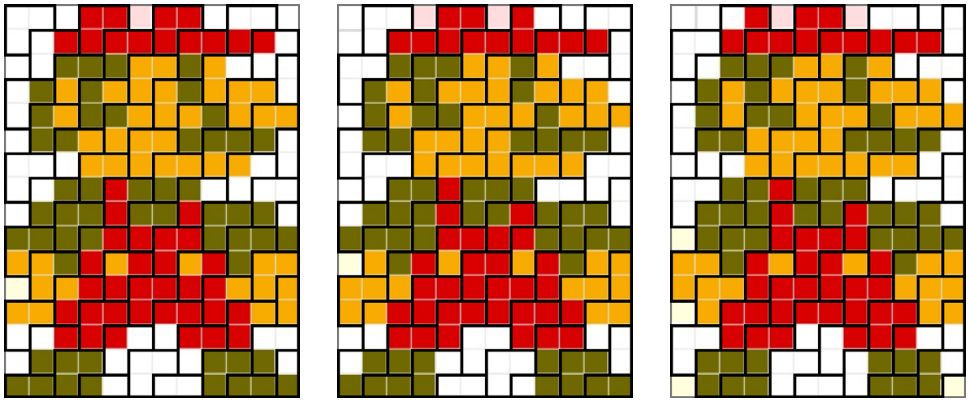
3.3. Rates of convergence of dual-based variables decomposition methods

Based on the primal–dual relation (Theorem 3.1), we can now convert the dual rates of convergence presented in the table of Section 3.2 to rates of convergence of the primal sequence. The new efficiency estimates for the primal sequence are given in the table below.

Method	Complexity result	Remarks
DBPG-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2C_1}{\sigma(k+1)}$	Ψ_i^* indicators, general m
DBPG-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2C_2}{\sigma(k+1)}$	General Ψ_i and m
DBPG-R	$\mathbb{E}(\|\mathbf{x}^k - \mathbf{x}^*\|^2) \leq \frac{2mC_3}{\sigma(m+k)}$	General Ψ_i and m
DAM-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2C_4}{\sigma k}$	$m = 2$
DAM-C	$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2C_5}{\sigma(k+1)}$	General Ψ_i and m

4. Numerical study

In this section we will illustrate the effectiveness and efficiency of the proposed algorithms for solving the total variation (TV) image denoising problem. The MATLAB code together with additional information regarding the implementation and the experiments could be found in the Dual Block Coordinate-Total Variation (DBC-TV) package that is available in the link http://tx.technion.ac.il/~yakovv/packages/DBC-TVpackage.zip.



(a) $\mathrm{TV}_I : \psi_1$. (b) $\mathrm{TV}_I : \psi_2$. (c) $\mathrm{TV}_I : \psi_3$.

Fig. 1. The decomposition of 16×12 pixels Mario image according to the isotropic TV into three separable functions. The images are partitioned into blocks of three pixels positioned in an r-shaped structure. Each block encompasses the three pixels that form the term $\sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2}$. Summing over all the terms represented by the blocks of any of the above images yields the appropriate separable function.

4.1. Total variation

The discrete TV ROF model for image denoising is given by the optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^{m \times n}} 0.5\|\mathbf{x} - \mathbf{b}\|_F^2 + \theta \cdot \mathrm{TV}_I(\mathbf{x}), \tag{TV}$$

where $\|\cdot\|_F$ stands for the Frobenius norm, $\mathbf{b}$ is the observed noisy image, $\mathbf{x}$ is the "true" image to be recovered and $\theta > 0$ is the regularization parameter that defines the trade-off between the data fidelity and regularity terms. The last component of the model is the semi-norm $\mathrm{TV}_I(\cdot)$ that stands for the discrete isotropic TV.

$$\mathbf{x} \in \mathbb{R}^{m \times n} \mathrm{TV}_I(\mathbf{x}) = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2} + \sum_{i=1}^{m-1} |x_{i,n} - x_{i+1,n}| + \sum_{j=1}^{n-1} |x_{m,j} - x_{m,j+1}|. \tag{4.1}$$

4.2. The decomposition

The expression (4.1) that defines the isotropic TV can be decomposed into separable functions. Under such decomposition, the model given by (TV) can be equivalently presented as the model (P). To describe the decomposition, we introduce the following notation. Let D_k denote the set of indices that correspond to the elements of the kth diagonal, where D_0 represents the indices set of the main diagonal, and D_k for $k > 0$ and $k < 0$ stands for the diagonals above and below the main diagonal, respectively. In addition, consider the partition of the diagonal indices set, $\{-(m-1), \ldots, n-1\}$, into three sets

$$K_i \equiv \left\{k \in \{-(m-1), \ldots, n-1\} : (k + 1 - i) \bmod 3 = 0\right\} \quad i = 1, 2, 3.$$

Now we are ready to write the definition of TV_I as

$$\begin{aligned} \mathrm{TV}_I(\mathbf{x}) &= \sum_{i=1}^{m} \sum_{j=1}^{n} \sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2} \\ &= \sum_{k \in K_1} \sum_{(i,j) \in D_k} \sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2} \\ &\quad + \sum_{k \in K_2} \sum_{(i,j) \in D_k} \sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2} \\ &\quad + \sum_{k \in K_3} \sum_{(i,j) \in D_k} \sqrt{(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2} \\ &= \psi_1(\mathbf{x}) + \psi_2(\mathbf{x}) + \psi_3(\mathbf{x}). \end{aligned}$$

This particular partition of the diagonals is the one with the smallest amount of sets such that for any $a, b \in K_i$ we have $|a - b| > 2$. The latter property guarantees that the functions g_i are separable. See the illustration in Fig. 1. We note that the decomposition considered here is very much different than the one considered in [8], where anisotropic or approximate-isotropic TV functions were considered.

Under the latter decomposition the optimization problems that need to be solved in the DAM and DBPG methods are separable and consequently easy to solve. Indeed, each such problem can be solved by solving simultaneously several 3-dimensional problems that can be solved by Newton's method employed on the dual problem.

4.3. Numerical results

We consider the isotropic TV denoising associated with the 512×512 "boat" image, available from the USC-SIPI Image Database—http://sipi.usc.edu/database/, that was contaminated by additive Gaussian noise with zero mean and standard deviation of 0.05. The numerical study was conducted on a PC with a 3.40 GHz processor with 8 GB RAM. It is easy to see that since f is a translation and scaling of the squared Euclidean function, DAM and DBPG are identical for the TV denoising problem. Thus, we will consider only DAM-C in our numerical study. The alternative algorithms that we consider are the alternating directions method of multipliers (ADMM) (see the review paper [7] and reference therein) and the dual FISTA method as proposed in [4]. Table 1 presents the number of iterations required by each algorithm in order to obtain an approximate optimality gap within a specified tolerance. Due to the fact that an optimal solution is not available, in order to compute an optimality gap, we considered an approximate solution, which we obtained by running dual FISTA for 10,000 iterations.

We can see that with respect to the convergence rate, DAM-C outperforms ADMM for the regularization parameter values under consideration except for large tolerance parameters with $\theta = 0.5$. In addition, DAM-C outperforms dual FISTA for small values of θ, but as the regularization parameter is increased, DAM-C is superior to dual FISTA in the first iterations but eventually, dual FISTA obtained better results. The per iteration complexity of ADMM is scalable to the one of DAM-C; hence, it is reasonable to compare

Table 1
Number of iterations required to obtain a solution with a specified optimality gap.

Tolerance	$\theta = 0.05$			$\theta = 0.1$			$\theta = 0.5$		
	DAM-C	ADMM	FISTA	DAM-C	ADMM	FISTA	DAM-C	ADMM	FISTA
$15 \cdot 10^{-2}$	2	4	3	3	5	6	25	23	40
$5 \cdot 10^{-2}$	3	7	7	7	10	16	93	87	103
$5 \cdot 10^{-3}$	15	27	28	50	81	67	725	896	336
$1 \cdot 10^{-3}$	37	>1000	58	122	>1000	133	>1000	>1000	610

Table 2
Number of iterations and time required for DAM-C and dual FISTA to obtain similar primal objective value ($\theta = 0.1$). The left table corresponds to exact minimization while the right table corresponds to an approximation via a single step of Newton's method.

Exact minimization				Approximate minimization			
DAM-C		FISTA		DAM-C		FISTA	
Iterations	Time (s)	Iterations	Time (s)	Iterations	Time (s)	Iterations	Time (s)
5	0.665	10	0.683	5	0.450	10	0.683
6	0.844	12	0.820	20	2.209	38	2.604
20	3.608	36	2.466	40	4.665	40	4.321
50	9.961	66	4.527	60	7.149	93	6.386
100	20.741	112	7.693	80	9.637	136	9.344
230	49.070	829	57.030	100	12.122	726	49.939
233	49.734	>10,000	>650	101	12.247	>10,000	>650

these algorithms based on the convergence rates of the primal objective function. However, to make a fair comparison between DAM-C and dual FISTA, we will compare in Table 2 the time that is required by the methods to obtain the same optimality gap.

We can see that, with respect to the running time, DAM-C outperforms dual FISTA in the first few iterations. In addition, after some point dual FISTA does not seem to significantly improve the primal objective value—a complication that does not occur for DAM-C. Inexact minimization of the dual of the problem that is solved at each iteration of DAM-C further enhances the aforementioned advantages of DAM-C, not only with respect to the execution time, but also with respect to the empirical convergence rate.

Acknowledgment

The work of Amir Beck was partially supported by the Israel Science Foundation grant #253/12.

References

[1] A. Auslender, Optimisation, in: Méthodes Numériques, Maîtrise de Mathématiques et Applications Fondamentales, Masson, Paris, 1976.
[2] A. Beck, On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes, SIAM J. Optim. 25 (1) (2015) 185–209.
[3] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (1) (2009) 183–202.
[4] A. Beck, M. Teboulle, A fast dual proximal gradient algorithm for convex minimization and applications, Oper. Res. Lett. 42 (1) (2014) 1–6.
[5] A. Beck, L. Tetruashvili, On the convergence of block coordinate descent type methods, SIAM J. Optim. 23 (2) (2013) 2037–2060.
[6] D.P. Bertsekas, J.N. Tsitsiklis, Parallel and Distributed Computation, Prentice-Hall International Editions, Englewood Cliffs, NJ, 1989.
[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.
[8] A. Chambolle, T. Pock, A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions, 2015. Preprint.
[9] P.L. Combettes, V.R. Wajs, Signal recovery by proximal forward–backward splitting, Multiscale Model. Simul. 4 (4) (2005) 1168–1200.
[10] L. Grippo, M. Sciandrone, Globally convergent block-coordinate techniques for unconstrained optimization, Optim. Methods Softw. 10 (1999) 587–637.
[11] M. Hong, X. Wang, M. Razaviyayn, Z.Q. Luo, Iteration complexity analysis of block coordinate decsnet methods, 2015. Preprint.
[12] Z. Lu, L. Xiao, On the complexity analysis of randomized block-coordinate descent methods, 2013. Preprint.
[13] T. Luo, P. Tseng, Error bounds and convergence analysis of feasible descent methods: a general approach, Ann. Oper. Res. 46 (1993) 157–178.
[14] J.J. Moreau, Proximitéet dualitédans un espace hilbertien, Bull. Soc. Math. France 93 (1965) 273–299.
[15] Y. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems, SIAM J. Optim. 22 (2) (2012) 341–362.
[16] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.
[17] P. Richtarik, M. Takac, Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function, Math. Program. A (2012).
[18] R.T. Rockafellar, Convex Analysis, Princeton Univ. Press, Princeton NJ, 1970.
[19] R.T. Rockafellar, R.J.B. Wets, Variational Analysis, in: Grundlehren der Mathematischen Wissenschaften, vol. 317, Springer-Verlag, Berlin, 1998.