A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions.

Antonin Chambolle, Thomas Pock

▶ **To cite this version:**

Antonin Chambolle, Thomas Pock. A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions.. 2014. hal-01099182v1

HAL Id: hal-01099182

https://hal.science/hal-01099182v1

Preprint submitted on 31 Dec 2014 (v1), last revised 18 Jun 2019 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions

Antonin Chambolle* and Thomas Pock†

December 30, 2014

Abstract

We analyze alternating descent algorithms for minimizing the sum of a quadratic function and block separable non-smooth functions. In case the quadratic interactions between the blocks are pairwise, we show that the schemes can be accelerated, leading to improved convergence rates with respect to related accelerated parallel proximal descent. As an application we obtain very fast algorithms for computing the proximity opertor of the 2D and 3D total variation.

1 Introduction

We discuss the acceleration of alternating minimization algorithms, for problems of the form

$$\min_{x=(x_i)_{i=1}^n} \mathcal{E}(x) := \sum_{i=1}^n f_i(x_i) + \frac{1}{2}\Big|\sum_{i=1}^n A_i x_i\Big|^2 \tag{1}$$

where each x_i lives in a Hilbert space $\mathcal{X}_i$, f_i are "simple" convex lsc functions whose proximity operator $(I+\tau\partial f_i)^{-1}$ can be easily evaluated, and the A_i are bounded linear operators from $\mathcal{X}_i$ to a common Hilbert space $\mathcal{X}$.

In general, we can check that for $n \geq 2$, alternating minimizations or descent methods do converge with rate $O(1/k)$ (where k is the number of iterations, see also [8]), and can hardly be accelerated. This is bad news, since clearly such a problem can be tackled by classical accelerated algorithms such as proposed in [30, 32, 31, 7, 38], yielding a $O(1/k^2)$ convergence rate for the objective. On the other hand, for $n = 2$ and $A_1 = A_2 = I_{\mathcal{X}}$ (the identity), we observe that alternating minimizations are nothing but a particular case of "forward-backward" descent, which can be accelerated by the above-mentioned methods.

*CMAP, Ecole Polytechnique, CNRS, 91128 Palaiseau, France.
e-mail: antonin.chambolle@cmap.polytechnique.fr

†Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria.
Digital Safety & Security Department, AIT Austrian Institute of Technology GmbH, 1220 Vienna, Austria.
e-mail: pock@icg.tugraz.at

Beyond these observations, our contribution is to analyze the descent properties of alternating minimizations (or implicit/explicit gradient steps as in [4, 39, 9], which might be simpler to perform), and in particular to show that acceleration is possible for general A_1, A_2 when $n = 2$. We also exhibit a structure which makes acceleration possible even for more than two variables. In these cases, the improvement with respect to a straight descent on the initial problem is essentially on the constant in the rate of convergence, since the Lipschitz constants in the partial descents are always smaller than the global Lipschitz constant of the smooth term $|\sum_i A_i x_i|^2$.

Problems of the form (1) arise in particular when applying Dykstra's algorithm to project on an intersection of (simple) convex sets, or more generally when computing the proximity operator

$$\min_z \sum_{i=1}^n g_i(z) + \frac{1}{2}|z - z^0|^2 \tag{2}$$

of a sum of simple convex functions $g_i(z)$ (for Dykstra's algorithm, the g_i's are the characteristic functions of convex sets).

Clearly, a dual problem to (2) can be written as (minus)

$$\min_{x=(x_i)_i} \sum_{i=1}^n (g_i^*(x_i) - \langle x_i, z^0 \rangle) + \frac{1}{2}\Big|\sum_{i=1}^n x_i\Big|^2 \tag{3}$$

which has exactly form (4) with $f_i(x_i) = g_i^*(x_i) - \langle x_i, z^0 \rangle$, and $A_i = I_{\mathcal{X}}$ for all i:

$$\min_{x=(x_i)_{i=1}^n} \sum_{i=1}^n f_i(x_i) + \frac{1}{2}\Big|\sum_{i=1}^n x_i\Big|^2, \tag{4}$$

Dysktra's algorithm is precisely an alternating minimization method for solving (4). Then, z is recovered by letting $z = z^0 - \sum_i x_i$.

Alternating minimization schemes for (1) (and more general problems) are widely found in the literature, as extensions of the linear Gauss-Seidel method. Many convergence results have been established, see in particular [4, 19, 36]. Our main results are valid for linearized alternating proximal descents, for which [4, 2, 39, 9] have provided convergence results (and rates when Kurdyka-Łojasiewicz inequalities [1] are available in [39]). See also [3, 11].

Two series of quite recent works are very close to our study. He, Yuan and collaborators [21, 18] have issued a series of papers where they tackle precisely the minimization of the same kind of energies, as a step in a more global Alternating Directions Method of Multipliers (ADMM) for energies involving more than two blocks. They could show a $O(1/k)$ rate of convergence of a measure of optimality for two classes of methods, one which consists into grouping the blocks into two subsets (which boils down then to the classical ADMM), another which consists in updating the step with a "Gaussian back substitution" after the descent steps. While some of their inequalities are very similar to ours, it does not seem that they give any new insight on acceleration strategies for (1).

On the other hand, two papers of Beck and Beck, Tetruashvili [8, 6] address rates of convergence for alternating descent algorithms, showing in a few cases a $O(1/k)$ decrease of the objective (and $O(1/k^2)$ for some smooth problems). We could show, adapting these approaches, that the same rate holds for the alternating minimization or proximal descent schemes for (1)

(which do not a priori enter the framework of [8, 6]). In addition, we exhibit a few situations where acceleration is possible, using the relaxation trick introduced in the FISTA algorithm [7] (see also [20, 30]). Unfortunately, these situations are essentially cases where the variable can be split into two sets of almost independent variables, limiting the number of interesting cases. We describe however in our last section that quite interesting examples can be solved with this technique, leading to a dramatic improvement with respect to standard approaches.

Eventually, stochastic alternating descents methods have been studied by many authors, in particular to deal with problems where the full gradient is too complicated to evaluate. First order methods with acceleration are discussed in [33, 28, 27, 17]. Some of these methods achieve very good convergence properties, however the proofs in these papers do not shed much light on the behaviour of deterministic methods, as the averaging typically will get rid of the antisymmetric terms which create difficulties in the proofs (and, up to now, prevent acceleration for general problems).

The plan of this paper is as follows: in the next section we recall the standard descent rule for the forward-backward splitting scheme (see for instance [15]), and show how it yields an accelerated rate of convergence for the FISTA overrelaxation. We also show that this is exactly equivalent to the alternating minimization of problems of form (4) with $n = 2$ variables.

Then, in the following section, we introduce the linearized alternating proximal descent scheme for (1) (which is also found in [39, 9], for more general problems). We show a rate of convergence for this descent, and exhibit cases which can be accelerated.

In the last section we illustrate these schemes with examples of possible splitting for solving the proximity operator of the Total Variation in 2 or 3 dimensions. This is related to recent domain decomposition approaches for this problem, see for instance [23]

2 The descent rule for Forward-Backward splitting and its consequences

2.1 The descent rule

Consider the standard problem

$$\min_{x\in\mathcal{X}} f(x) + g(x) \tag{5}$$

where f is a proper convex lsc function and g is a $C^{1,1}$ convex function, whose gradient has Lipschitz constant L, both defined on a Hilbert space $\mathcal{X}$. We can define the simple Forward-Backward descent scheme as the iteration of the operator:

$$T\bar{x} = \hat{x} := (I + \tau\partial f)^{-1}(I - \tau\nabla g)(\bar{x}). \tag{6}$$

Then, it is well-known [31, 7] that the objective $F(x) = f(x)+g(x)$ satisfies the following descent rule:

$$F(x) + \frac{1}{2\tau}|x - \bar{x}|^2 \geq F(\hat{x}) + \frac{1}{2\tau}|x - \hat{x}|^2 \tag{7}$$

as soon as $\tau \leq 1/L$. An elementary way to show this is to observe that $\hat{x}$ is the minimizer of the strongly convex function $f(x) + \langle \nabla g(\bar{x}), x\rangle + |x-\bar{x}|^2/(2\tau)$, so that

$$\begin{aligned} f(x) + g(x) + \frac{|x-\bar{x}|^2}{2\tau} &\geq f(x) + g(\bar{x}) + \langle \nabla g(\bar{x}), x - \bar{x}\rangle + \frac{|x-\bar{x}|^2}{2\tau} \\ &\geq f(\hat{x}) + g(\bar{x}) + \langle \nabla g(\bar{x}), \hat{x} - \bar{x}\rangle + \frac{|\hat{x}-\bar{x}|^2}{2\tau} + \frac{|x-\hat{x}|^2}{2\tau} \\ &\geq f(\hat{x}) + g(\hat{x}) + \left(\frac{1}{\tau} - L\right)\frac{|\hat{x}-\bar{x}|^2}{2} + \frac{|x-\hat{x}|^2}{2\tau} \end{aligned}$$

where in the last line, we have used the fact that ∇g is L-Lipschitz.

One derives easily that the standard Forward-Backward descent scheme (which consists in letting $x^{k+1} = Tx^k$) converges with rate $O(1/k)$, and precisely that if x^* is a minimizer,

$$F(x^k) - F(x^*) \leq L\frac{|x^* - x^0|^2}{2k} \tag{8}$$

if $\tau = 1/L$.

2.2 Acceleration

However, it is easy to derive from (7) a faster descent, as shown in [7] (see also [30, 31, 32, 38]). Indeed, letting in (7)

$$\hat{x} = x^{k+1}, \quad \bar{x} = x^k + \frac{t_k - 1}{t_{k+1}}(x^k - x^{k-1}), \quad x = \frac{(t_{k+1}-1)x^k + x^*}{t_{k+1}}$$

(and $x^{-1} := x^0$) and rearranging, for a given sequence of real numbers $t_k \geq 1$, we obtain

$$\begin{aligned} &\frac{|(t_{k+1}-1)x^k + x^* - t_{k+1}x^{k+1}|^2}{2\tau} + t_{k+1}^2(F(x^{k+1}) - F(x^*)) \\ &\qquad \leq \frac{|(t_k-1)x^{k-1} + x^* - t_k x^k|^2}{2\tau} + (t_{k+1}^2 - t_{k+1})(F(x^k) - F(x^*)) \end{aligned}$$

If $t_{k+1}^2 - t_{k+1} \leq t_k^2$, this can easily iterated. For instance, letting $t_k = (k+1)/2$[1] and $\tau = 1/L$, one finds

$$F(x^k) - F(x^*) \leq 2L\frac{|x^* - x^0|^2}{(k+1)^2} \tag{9}$$

which is nearly optimal in regards to the lower bounds shown in [29, 31].

What is interesting to observe, here, is that this acceleration property will be true for any scheme for which a descent rule such as (7) holds, with the same proof. We will now show that this is also the case for an alternating descent of the form (4), if $n = 2$.

[1] Other choices might yield better properties, see in particular [14].

2.3 The descent rule for two-variables alternating descent

Consider now problem (4), with $n = 2$:

$$\min_{x=(x_1,x_2)\in\mathcal{X}^2} \mathcal{E}_2(x) := f_1(x_1) + f_2(x_2) + \frac{1}{2}|x_1 + x_2|^2. \tag{10}$$

We consider the following algorithm, which transforms $\bar{x}$ into $T\bar{x} = \hat{x}$ by letting

$$\hat{x}_1 := \arg\min_{x_1} f_1(x_1) + \frac{1}{2}|x_1 + \bar{x}_2|^2 \tag{11}$$

$$\hat{x}_2 := \arg\min_{x_2} f_2(x_2) + \frac{1}{2}|\hat{x}_1 + x_2|^2. \tag{12}$$

Then by strong convexity, for any $x_1 \in \mathcal{X}$, we have

$$f_1(x_1) + \frac{1}{2}|x_1 + \bar{x}_2|^2 \geq f_1(\hat{x}_1) + \frac{1}{2}|\hat{x}_1 + \bar{x}_2|^2 + \frac{1}{2}|x_1 - \hat{x}_1|^2$$

while for any $x_2 \in \mathcal{X}$,

$$f_2(x_2) + \frac{1}{2}|\hat{x}_1 + x_2|^2 \geq f_2(\hat{x}_2) + \frac{1}{2}|\hat{x}_1 + \hat{x}_2|^2 + \frac{1}{2}|x_2 - \hat{x}_2|^2.$$

Summing these inequalities, we find

$$\begin{aligned} f_1(x_1) + f_2(x_2) + \frac{1}{2}|x_1 + x_2|^2 \geq \\ f_1(\hat{x}_1) + f_2(\hat{x}_2) + \frac{1}{2}|\hat{x}_1 + \hat{x}_2|^2 + \frac{1}{2}|x_1 - \hat{x}_1|^2 + \frac{1}{2}|x_2 - \hat{x}_2|^2 \\ + \frac{1}{2}\left(|x_1 + x_2|^2 - |x_1 + \bar{x}_2|^2 + |\hat{x}_1 + \bar{x}_2|^2 - |\hat{x}_1 + x_2|^2\right). \end{aligned}$$

Now, the last line in this equation is

$$(x_1 - \hat{x}_1)(x_2 - \bar{x}_2) = \frac{1}{2}|x_1 + x_2 - (\hat{x}_1 + \bar{x}_2)|^2 - \frac{1}{2}|x_1 - \hat{x}_1|^2 - \frac{1}{2}|x_2 - \bar{x}_2|^2$$

and it follows

$$\begin{aligned} f_1(x_1) + f_2(x_2) + \frac{1}{2}|x_1 + x_2|^2 + \frac{1}{2}|x_2 - \bar{x}_2|^2 \\ \geq f_1(\hat{x}_1) + f_2(\hat{x}_2) + \frac{1}{2}|\hat{x}_1 + \hat{x}_2|^2 + \frac{1}{2}|x_2 - \hat{x}_2|^2. \end{aligned} \tag{13}$$

As before, one easily deduces the following result:

Proposition 1. *Let $x^0 = x^{-1}$ be given and for each k let $\bar{x}_2^k = x_2^k + \frac{k-1}{k+2}(x_2^k - x_2^{k-1})$, x_1^{k+1} minimize $\mathcal{E}_2(\cdot, \bar{x}_2^k)$ and x_2^{k+1} minimize $\mathcal{E}_2(x_1^{k+1}, \cdot)$. Call x^* a global minimizer of $\mathcal{E}_2$. Then*

$$\mathcal{E}_2(x^k) - \mathcal{E}_2(x^*) \leq 2\frac{|x_2^* - x_2^0|^2}{(k+1)^2}. \tag{14}$$

However, we have to precise here that this result is *the same* as the main result in [7] recalled in the previous section, for elementary reasons which we explain in the next Section 2.4. Observe on the other hand that a straight application of [7] to problem (10) (with $|x_1 + x_2|^2/2$ as the smooth term), that is, governed by the iteration

$$\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \left(I + \tau \begin{pmatrix} \partial f_1 \\ \partial f_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} \bar{x}_1 - \tau(\bar{x}_1 + \bar{x}_2) \\ \bar{x}_2 - \tau(\bar{x}_1 + \bar{x}_2) \end{pmatrix},$$

needs $\tau \leq 1/2$ and hence yields the estimate

$$\mathcal{E}_2(x^k) - \mathcal{E}_2(x^*) \leq 4 \frac{|x_1^* - x_1^0|^2 + |x_2^* - x_2^0|^2}{(k+1)^2}.$$

which is less good than (14), whereas the parallel algorithm in [17], with a deterministic block (x_1, x_2), is achieving the same rate.

2.4 The two previous problems are identical

In fact, what is not obvious at first glance is that problem (5) and (10) are exactly identical, while the Forward-Backward descent scheme for the first is the same as the alternating minimization for the second. Indeed, g has L-Lipschitz gradient if and only if there exists a convex function g^0 such that

$$g(x) = \min_{y \in \mathcal{X}} g^0(y) + L \frac{|x-y|}{2}^2. \tag{15}$$

(An elementary way to show this is to consider that g^*, the Legendre-Fenchel transform of g, is $(1/L)$-strongly convex, which means that $g^* - (1/L)|z|^2/2$ is convex. The function g^0 is then obtained as the Legendre-Fenchel transform of the latter.) It is also standard (see for instance [12]) that the minimizer y in (15) is nothing else as the point $x - (1/L)\nabla g(x)$.

Hence, (5) can be rewritten as the minimization problem

$$\min_{x,y \in \mathcal{X}} f(x) + g^0(y) + L \frac{|x-y|}{2}^2. \tag{16}$$

and the Forward-Backward scheme (6) with step $\tau = 1/L$ is an alternating descent scheme for problem (16), first minimizing in y and then in x.

Observe that the descent rule (13) is a bit more precise, though, than (7). In particular, it also implies the same accelerated rate for the scheme (minimizing first in x, and then in y):

$$x^{k+1} = (I + \tau \partial f)^{-1} \left(y^k + \tfrac{t_k - 1}{t_{k+1}} (y^k - y^{k-1}) \right) \tag{17}$$

$$y^{k+1} = x^{k+1} - \tau \nabla g(x^{k+1}) \tag{18}$$

with t_k as before, which coincides with [7] only when ∇g is linear.

3 Proximal alternating descent

Now we turn to problem (1) which is a bit more general, and less trivially reduced to another standard problem as in the previous section. In general, we can not always assume that it is

possible to exactly minimize (1) with respect to one variable x_i. However, we can perform a gradient descent on the variable x_i by solving problems of the form

$$\min_{x_i} f_i(x_i) + \left\langle A_i x_i, \sum_j A_j \bar{x}_j \right\rangle + \frac{\langle M_i(x_i - \bar{x}_i), x_i - \bar{x}_i \rangle}{2\tau_i}$$

for any given $\bar{x}$, where M_i is a nonnegative operator defining a metric for the variable x_i as soon as f_i is "simple" enough in the given metrics. Then, in case M_i/τ_i is precisely $A_i^* A_i$, the solution of this problem is a minimizer of

$$\min_{x_i} f_i(x_i) + \frac{1}{2}\Big|A_i x_i + \sum_{j \neq i} A_j \bar{x}_j\Big|^2,$$

so that the alternating minimization algorithm can be considered as a special case of the alternating descent algorithm (which will require only $M_i/\tau_i \geq A_i^* A_i$). In the sequel to simplify we will consider the standard metric corresponding to $M_i = I_{\mathcal{X}_i}$, however any other metric for which the prox of f_i could be calculated is admissible in practice. Hence we will focus on alternating descent steps for problem (1) in the standard metrics. The alternating proximal scheme seems to be first found in [4, Alg. 4.1], while the linearized version we are focusing is proposed and studied in [39, 9].

This can be described as follows: we let for each i, $\tau_i > 0$, and we produce $\hat{x}$ by minimizing

$$\min_{x_i} f_i(x_i) + \left\langle x_i, A_i^*\Big(\sum_{j<i} A_j \hat{x}_j + \sum_{j \geq i} A_j \bar{x}_j\Big)\right\rangle + \frac{|x_i - \bar{x}_i|^2}{2\tau_i}.$$

Now, for all x_i,

$$\begin{aligned}
f_i(x_i) + \frac{1}{2}\Big|A_i x_i + \sum_{j<i} A_j \hat{x}_j + \sum_{j>i} A_j \bar{x}_j\Big|^2 + \frac{|x_i - \bar{x}_i|^2}{2\tau_i} &= \\
f_i(x_i) + \frac{1}{2}\Big|\sum_{j<i} A_j \hat{x}_j + \sum_{j \geq i} A_j \bar{x}_j\Big|^2 + \left\langle x_i - \bar{x}_i, A_i^*\Big(\sum_{j<i} A_j \hat{x}_j + \sum_{j \geq i} A_j \bar{x}_j\Big)\right\rangle & \\
+ \frac{1}{2}\Big|A_i(x_i - \bar{x}_i)\Big|^2 + \frac{|x_i - \bar{x}_i|^2}{2\tau_i} & \\
\geq f_i(\hat{x}_i) + \frac{1}{2}\Big|\sum_{j<i} A_j \hat{x}_j + \sum_{j \geq i} A_j \bar{x}_j\Big|^2 + \left\langle \hat{x}_i - \bar{x}_i, A_i^*\Big(\sum_{j<i} A_j \hat{x}_j + \sum_{j \geq i} A_j \bar{x}_j\Big)\right\rangle & \\
+ \frac{1}{2}\Big|A_i(x_i - \bar{x}_i)\Big|^2 + \frac{|\hat{x}_i - \bar{x}_i|^2}{2\tau_i} + \frac{|x_i - \hat{x}_i|^2}{2\tau_i} & \\
= f_i(\hat{x}_i) + \frac{1}{2}\Big|\sum_{j \leq i} A_j \hat{x}_j + \sum_{j>i} A_j \bar{x}_j\Big|^2 - \frac{1}{2}\Big|A_i(\hat{x}_i - \bar{x}_i)\Big|^2 & \\
+ \frac{1}{2}\Big|A_i(x_i - \bar{x}_i)\Big|^2 + \frac{|\hat{x}_i - \bar{x}_i|^2}{2\tau_i} + \frac{|x_i - \hat{x}_i|^2}{2\tau_i} &
\end{aligned}$$

Letting for each i, $B_i = (1/\tau_i)I - A_i^* A_i$ and assuming $B_i \geq 0$, it follows

$$f_i(x_i) + \frac{1}{2}\Big|A_i x_i + \sum_{j<i} A_j \hat{x}_j + \sum_{j>i} A_j \bar{x}_j\Big|^2 + \frac{|x_i - \bar{x}_i|^2_{B_i}}{2} \geq$$
$$f_i(\hat{x}_i) + \frac{1}{2}\Big|\sum_{j\leq i} A_j \hat{x}_j + \sum_{j>i} A_j \bar{x}_j\Big|^2 + \frac{|\hat{x}_i - \bar{x}_i|^2_{B_i}}{2} + \frac{|x_i - \hat{x}_i|^2}{2\tau_i} \quad (19)$$

Summing over all i, we find:

$$\mathcal{E}(x) + \frac{\|x - \bar{x}\|^2_B}{2} \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x} - \bar{x}\|^2_B}{2} + \sum_{i=1}^n \frac{|x_i - \hat{x}_i|^2}{2\tau_i}$$
$$+ \frac{1}{2}\Bigg(\Big|\sum_{j=1}^n A_j x_j\Big|^2 - \Big|\sum_{j=1}^n A_j \hat{x}_j\Big|^2$$
$$+ \sum_{i=1}^n \Bigg(\Big|\sum_{j\leq i} A_j \hat{x}_j + \sum_{j>i} A_j \bar{x}_j\Big|^2 - \Big|A_i x_i + \sum_{j<i} A_j \hat{x}_j + \sum_{j>i} A_j \bar{x}_j\Big|^2\Bigg)\Bigg) \quad (20)$$

where $\|x\|^2_B = \sum_{i=1}^n |x_i|^2_{B_i}$. Denoting $y_i = A_i x_i$ we can rewrite the last two lines of this formula (with obvious notation) as follows:

$$-\frac{1}{2}\Big|\sum_{i=1}^n y_i - \hat{y}_i\Big|^2 + \sum_{i=1}^n \Big\langle y_i - \hat{y}_i, \textstyle\sum_j y_j\Big\rangle + \sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \sum_{j<i}\hat{y}_j + \frac{y_i + \hat{y}_i}{2} + \sum_{j>i}\bar{y}_j\Big\rangle$$
$$= -\frac{1}{2}\Big|\sum_{i=1}^n y_i - \hat{y}_i\Big|^2 + \frac{1}{2}\sum_{i=1}^n |\hat{y}_i - y_i|^2$$
$$+ \sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \textstyle\sum_{j<i}(\hat{y}_j - y_j) + \sum_{j>i}(\bar{y}_j - y_j)\Big\rangle \quad (21)$$

Notice then that

$$\sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \textstyle\sum_{j<i}(\hat{y}_j - y_j)\Big\rangle = \sum_{j=1}^n \sum_{i>j} \langle \hat{y}_i - y_i, \hat{y}_j - y_j\rangle$$
$$= \sum_{i=1}^n \sum_{i<j} \langle \hat{y}_j - y_j, \hat{y}_i - y_i\rangle = \frac{1}{2}\sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \textstyle\sum_{j\neq i}(\hat{y}_j - y_j)\Big\rangle$$
$$= \frac{1}{2}\left(\Big|\sum_{i=1}^n \hat{y}_i - y_i\Big|^2 - \sum_{i=1}^n |\hat{y}_i - y_i|^2\right) \quad (22)$$

so that (21) boils down to

$$\sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \textstyle\sum_{j>i}(\bar{y}_j - y_j)\Big\rangle.$$

One deduces from (20) that for all x,

$$\mathcal{E}(x) + \frac{\|x-\bar{x}\|_B^2}{2} \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x}-\bar{x}\|_B^2}{2} + \sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i), \textstyle\sum_{j>i} A_j(\bar{x}_j - x_j)\Big\rangle + \sum_{i=1}^n \frac{|x_i-\hat{x}_i|^2}{2\tau_i}. \quad (23)$$

This can also be written

$$\mathcal{E}(x) + \frac{\|x-\bar{x}\|_B^2}{2} \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x}-\bar{x}\|_B^2}{2} + \frac{\|x-\hat{x}\|_B^2}{2} + \frac{1}{2}\sum_{i=1}^n |A_i(x_i-\hat{x}_i)|^2 + \sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i), \textstyle\sum_{j>i} A_j(\bar{x}_j - x_j)\Big\rangle. \quad (24)$$

Then, using (22) again, one has

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^n |A_i(x_i-\hat{x}_i)|^2 &+ \sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i), \textstyle\sum_{j>i} A_j(\bar{x}_j - x_j)\Big\rangle \\ &= \frac{1}{2}\sum_{i=1}^n |y_i-\hat{y}_i|^2 + \sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \textstyle\sum_{j>i}(\bar{y}_j - \hat{y}_j)\Big\rangle + \sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \textstyle\sum_{j>i}(\hat{y}_j - y_j)\Big\rangle \\ &\qquad \frac{1}{2}\Big|\sum_{i=1}^n y_i - \hat{y}_i\Big|^2 + \sum_{i=1}^n \Big\langle \hat{y}_i - y_i, \textstyle\sum_{j>i}(\bar{y}_j - \hat{y}_j)\Big\rangle.\end{aligned}$$

which, combined to (24), yields also the following estimate:

$$\mathcal{E}(x) + \frac{\|x-\bar{x}\|_B^2}{2} \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x}-\bar{x}\|_B^2}{2} + \frac{\|x-\hat{x}\|_B^2}{2} + \frac{1}{2}\Big|\sum_{i=1}^n A_i(x_i-\hat{x}_i)\Big|^2 + \sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i), \textstyle\sum_{j>i} A_j(\bar{x}_j - \hat{x}_j)\Big\rangle. \quad (25)$$

3.1 A $O(1/k)$ convergence rate

Convergence of the alternating proximal minimization scheme in this framework (and more general ones, see for instance [4]), in the sense that (x^k) is a minimizing sequence, is well-known and not so difficult to establish. In case the energy is coercive, we can obtain from (23) a $O(1/k)$ decay estimate after $k \times n$ alternating minimizations, following essentially the similar proofs in [8, 6]. The idea is first to consider $x = \bar{x}$ in (23), yielding

$$\mathcal{E}(\bar{x}) \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x}-\bar{x}\|_B^2}{2} + \sum_{i=1}^n \frac{|\hat{x}_i-\bar{x}_i|^2}{2\tau_i}.$$

In particular, if x^* is a solution, letting $\bar{x} = x^k$, it follows

$$\mathcal{E}(x^{k+1}) - \mathcal{E}(x^*) + \frac{\|x^{k+1}-x^k\|_B^2}{2} + \sum_{i=1}^n \frac{|x_i^{k+1}-x_i^k|^2}{2\tau_i} \leq \mathcal{E}(x^k) - \mathcal{E}(x^*) \quad (26)$$

A rate will follow if we can show that $\|\hat{x} - \bar{x}\|$ bounds $\mathcal{E}(\hat{x}) - \mathcal{E}(x^*)$. From (25) we obtain, choosing $x = x^*$,

$$\mathcal{E}(\hat{x}) - \mathcal{E}(x^*) + \frac{\|\hat{x} - \bar{x}\|_B^2}{2} + \frac{\|x^* - \hat{x}\|_B^2}{2} + \frac{1}{2}\Big|\sum_{i=1}^n A_i(x_i^* - \hat{x}_i)\Big|^2 + \sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i^*), \textstyle\sum_{j>i} A_j(\bar{x}_j - \hat{x}_j)\Big\rangle \leq \frac{\|x^* - \bar{x}\|_B^2}{2}$$

Now, since

$$\frac{\|x^* - \bar{x}\|_B^2}{2} - \frac{\|\hat{x} - \bar{x}\|_B^2}{2} - \frac{\|x^* - \hat{x}\|_B^2}{2} = \langle \hat{x} - x^*, \bar{x} - \hat{x}\rangle_B$$

this is also

$$\mathcal{E}(\hat{x}) - \mathcal{E}(x^*) + \frac{1}{2}\Big|\sum_{i=1}^n A_i(x_i^* - \hat{x}_i)\Big|^2 \leq \langle \hat{x} - x^*, \bar{x} - \hat{x}\rangle_B - \sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i^*), \textstyle\sum_{j>i} A_j(\bar{x}_j - \hat{x}_j)\Big\rangle,$$

and there exists C (depending on the A_i's) such that

$$\mathcal{E}(x^{k+1}) - \mathcal{E}(x^*) \leq C\sqrt{\sum_{i=1}^n \frac{|x_i^{k+1} - x_i^k|^2}{2\tau_i}}\, \|x^{k+1} - x^*\|.$$

Thus, (26) yields, letting $\Delta_k := \mathcal{E}(x^k) - \mathcal{E}(x^*)$,

$$\Delta_{k+1} + \frac{1}{C^2\|x^{k+1} - x^*\|^2}\Delta_{k+1}^2 \leq \Delta_k. \tag{27}$$

It follows:

Proposition 2. *Assume that $\mathcal{E}$ is coercive. Then the alternating minimization algorithm produces a sequence (x^k) such that*

$$\mathcal{E}(x^k) - \min \mathcal{E} \leq O\left(\frac{1}{k}\right).$$

Proof. Indeed, if $\mathcal{E}$ is coercive, one has that $\|x^k - x^*\|$ is a bounded sequence and (27) reads

$$\Delta_{k+1} + \tilde{C}\Delta_{k+1}^2 \leq \Delta_k. \tag{28}$$

Then, it follows from [6, Lemma 3.6] that

$$\Delta_k \leq \max\left\{\frac{\Delta_0}{2^{\frac{k-1}{2}}}, \frac{4}{\tilde{C}(k-1)}\right\}.$$

For the reader's convenience, we give a variant of Amir Beck's proof, which shows a slightly different estimate (and that, in fact, asymptotically, the "4" can be reduced). We can let $x_k = \tilde{C}\Delta_k$, with this normalization we get that

$$x_{k+1}(1 + x_{k+1}) \leq x_k \;\Rightarrow\; x_{k+1} \leq \frac{-1 + \sqrt{1 + 4x_k}}{2},$$

and

$$x_k x_{k+1}^{-2} - x_{k+1}^{-1} - 1 \geq 0$$

so that

$$\frac{1}{x_{k+1}} \geq \frac{1+\sqrt{1+4x_k}}{2x_k} \geq \frac{1}{x_k} + 1 - x_k. \tag{29}$$

Notice that from the first relationship, we find that

$$x_{k+1} + \frac{1}{4} \leq x_{k+1} + \frac{1}{2} \leq \sqrt{x_k + \frac{1}{4}}$$

which yields

$$x_{k+1} \leq \left(x_0 + \frac{1}{4}\right)^{\frac{1}{2^{k+1}}} - \frac{1}{2}.$$

In particular it takes only

$$\bar{k} \geq \frac{\log\log(x_0 + 1/2) - \log\log 5/4}{\log 2}$$

iterations to reach $x_{\bar{k}} \leq 3/4$, which is for instance 7 iterations if $x_0 \approx 10^{20}$, and one more iteration to reach $x_{\bar{k}+1} \leq 1/2$. Then thanks to (29), one has for $k \geq \bar{k}+1$ that $1/x_{k+1} \geq 1/x_k + 1/2 \geq 1/x_{\bar{k}+1} + (k-\bar{k})/2$, yielding $\tilde{C}\Delta_k \leq (2+\epsilon)/k$ for any $\epsilon > 0$ and k large enough. Using (29) again one sees that this bound can, in fact, be made as close as wanted to $1/k$ (but for k large). □

Remark 1. One can observe that as x^k converges to the set of solutions (which will be true in finite dimension), then the constant $\tilde{C} \geq 1/(C^2\|x^{k+1} - x^*\|^2)$ in (28) is improving, yielding a better actual rate. In particular if $\mathcal{E}$ satisfies in addition a Kurdyka-Łojasiewicz type inequality [1] near a limiting point x^*, then this global rate should be improved.

3.2 The case $n = 2$

In case $n = 2$, the situation is simpler (as for the alternating minimizations, for which [8] already showed that acceleration is possible for smooth functions). Indeed, (23) shows that

$$\begin{aligned}
&\mathcal{E}(x) + \frac{\|x - \bar{x}\|_B^2}{2} \\
&\qquad \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x} - \bar{x}\|_B^2}{2} + \langle A_1(\hat{x}_1 - x_1), A_2(\bar{x}_2 - x_2)\rangle + \frac{1}{2}\sum_{i=1}^n \frac{|x_i - \hat{x}_i|^2}{2\tau_i} \\
&\qquad\qquad \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x} - \bar{x}\|_B^2}{2} - \frac{|A_1(\hat{x}_1 - x_1)|^2}{2} - \frac{|A_2(\bar{x}_2 - x_2)|^2}{2} + \frac{1}{2}\sum_{i=1}^n \frac{|x_i - \hat{x}_i|^2}{2\tau_i}
\end{aligned}$$

so that

$$\mathcal{E}(x) + \frac{|x_1 - \bar{x}_1|_{B_1}^2}{2} + \frac{|x_2 - \bar{x}_2|^2}{2\tau_2} \geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x} - \bar{x}\|_B^2}{2} + \frac{|x_1 - \hat{x}_1|_{B_1}^2}{2} + \frac{|x_2 - \hat{x}_2|^2}{2\tau_2} \tag{30}$$

This makes the FISTA acceleration of [7] possible for this scheme, yielding the rate

$$\mathcal{E}(x^k) - \mathcal{E}(x^*) \leq \frac{2}{(k+1)^2}\left(|x_1^0 - x_1^*|_{B_1}^2 + \frac{|x_2^0 - x_2^*|}{\tau_2}\right)$$

as explained in Section 2.2. If one can do an exact minimization with respect to x_1, one has in addition $B_1 = 0$ and falls back into the situation described in Section 2.3. Moreover, if one also performs exact minimizations with respect to x_2, then the rate becomes

$$\mathcal{E}(x^k) - \mathcal{E}(x^*) \leq 2\frac{|x_2^0 - x_2^*|_{A_2^* A_2}}{(k+1)^2}$$

which is the generalization of (14).

3.3 A more general case which can be accelerated

In fact, the case $n = 2$ is a particular case of a more general situation where the interaction term can be written as the sum of pairwise interactions between two variables x_i and x_j.

Formally, it means that for all i, j, there exists $A_{i,j}$ a bounded linear operator from $\mathcal{X}_i$ to a Hilbert space $\mathcal{X}_{i,j} = \mathcal{X}_{j,i}$ such that for all $x = (x_i)_{i=1}^n \in \times_{i=1}^n \mathcal{X}_i$,

$$\Big|\sum_{j=1}^n A_j x_j\Big|^2 = \sum_{1\leq i<j\leq n} |A_{i,j}x_i + A_{j,i}x_j|^2.$$

In this case, one checks that for any $(x_i, x_j) \in \mathcal{X}_i \times \mathcal{X}_j$, if $i < j$ then

$$\langle A_i x_i, A_j x_j\rangle = \langle A_{i,j}x_i, A_{j,i}x_j\rangle,$$

while for $i = 1, \ldots, n$,

$$|A_i x_i|^2 = \sum_{j\neq i} |A_{i,j}x_i|^2.$$

It follows

$$\begin{aligned}\sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i), \textstyle\sum_{j>i} A_j(\bar{x}_j - x_j)\Big\rangle &= \sum_{1\leq i<j\leq n} \langle A_{i,j}(\hat{x}_i - x_i), A_{j,i}(\bar{x}_j - x_j)\rangle \\ &\geq -\frac{1}{2}\sum_{1\leq i<j\leq n} \big(|A_{i,j}(x_i - \hat{x}_i)|^2 + |A_{j,i}(x_j - \bar{x}_j)|^2\big) \\ &= -\frac{1}{2}\sum_{i=1}^n \sum_{i<j\leq n} |A_{i,j}(x_i - \hat{x}_i)|^2 - \frac{1}{2}\sum_{i=1}^n \sum_{1\leq j<i} |A_{i,j}(x_i - \bar{x}_i)|^2,\end{aligned}$$

where in the last term we have exchanged the indices i and j. On the other hand,

$$\frac{1}{2}\sum_{i=1}^n |A_i(x_i - \hat{x}_i)|^2 = \frac{1}{2}\sum_{i=1}^n \sum_{j\neq i} |A_{i,j}(x_i - \hat{x}_i)|^2$$

and it follows that

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^n |A_i(x_i - \hat{x}_i)|^2 + \sum_{i=1}^n \Big\langle A_i(\hat{x}_i - x_i), \textstyle\sum_{j>i} A_j(\bar{x}_j - x_j)\Big\rangle&. \\ \geq \frac{1}{2}\sum_{i=1}^n \sum_{1\leq j<i} |A_{i,j}(x_i - \hat{x}_i)|^2 - \frac{1}{2}\sum_{i=1}^n \sum_{1\leq j<i} |A_{i,j}(x_i - \bar{x}_i)|^2&.\end{aligned}$$

We therefore deduce from (24) that

$$\mathcal{E}(x) + \frac{\|x-\bar{x}\|_B^2}{2} + \frac{1}{2}\sum_{i=1}^{n}\sum_{1\le j<i}|A_{i,j}(x_i-\bar{x}_i)|^2.$$
$$\geq \mathcal{E}(\hat{x}) + \frac{\|\hat{x}-\bar{x}\|_B^2}{2} + \frac{\|x-\hat{x}\|_B^2}{2} + \frac{1}{2}\sum_{i=1}^{n}\sum_{1\le j<i}|A_{i,j}(x_i-\hat{x}_i)|^2. \quad (31)$$

It follows, once more, that FISTA-like acceleration is possible for this alternating minimization strategy, yielding the rate

$$\mathcal{E}(x^k) - \mathcal{E}(x^*) \leq \frac{2}{(k+1)^2}\left(\frac{\|x^0-x^*\|_B^2}{2} + \frac{1}{2}\sum_{i=1}^{n}\sum_{1\le j<i}|A_{i,j}(x_i^0-x_i^*)|^2\right).$$

for any minimizer x^*.

4 Application: various splitting strategies for Total Variation minimization

In this section, we consider different splitting algorithms for minimizing the Rudin, Osher, Fatemi (ROF) model for total variation (TV)-based image denoising.

$$\min_u \mathrm{TV}_p(u) + \frac{\lambda}{2}\|u-f\|_2^2, \quad (32)$$

where $f \in \mathbb{R}^{MN}$ is the (noisy) input image and $\lambda > 0$ is a regularization parameter. TV_p corresponds to a discrete ℓ_p-norm ($p \in \{1,2\}$) based approximation of the total variation. We will denote by $\hat{u}$ the unique minimizer of (32). The exact definition of the total variation function TV_p will depend on the certain type of the splitting strategy and hence it will be detailed in the respective sections.

Let us fix some notation. An image x is defined on a $M \times N$ pixel grid with indices $(1,1) \le (i,j) \le (M,N)$ which is re-organized into a single column vector $u \in \mathbb{R}^{MN}$ but for the ease of notation we will keep the structure of the indices. We will also make use of the function

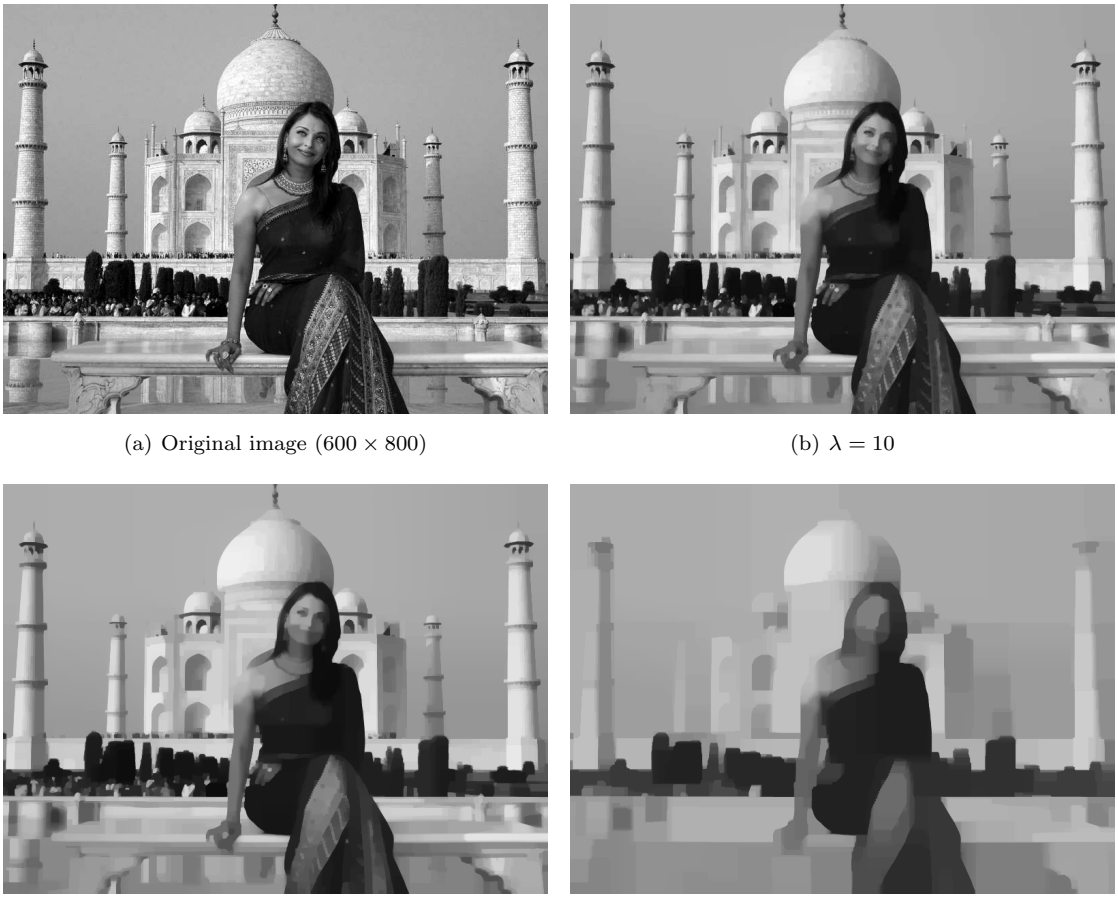
$$\delta_C(x) = \begin{cases} 0 & \text{if} \quad x \in C \\ \infty & \text{else} \end{cases}$$

which denotes the indicator function for a convex set C.

4.1 Chain-based splitting

In this section, we consider the anisotropic version of (32) ($p = 1$)

$$\min_u \mathcal{P}(u) = \mathrm{TV}_1(u) + \frac{\lambda}{2}\|u-f\|^2, \quad (33)$$



(a) Original image (600×800) (b) $\lambda = 10$

(c) $\lambda = 5$ (d) $\lambda = 1$

Figure 1: Test image of size 600×800 with intensity values in the range $[0, 1]$ used in our experiments. We consider experiments with different strength of the regularization parameter to study the behavior of the algorithm in these cases.

which allows a splitting of the total variation as $\mathrm{TV}_1(u) = \mathrm{TV}_h(u) + \mathrm{TV}_v(u)$, where

$$\mathrm{TV}_h(u) = \sum_{i,j=1}^{M,N-1} |u_{i,j+1} - u_{i,j}|$$

computes the total variation along the horizontal edges and

$$\mathrm{TV}_v(u) = \sum_{i,j=1}^{M-1,N} |u_{i+1,j} - u_{i,j}|$$

computes the total variation along the vertical edges. See Figure 2 for a simple example, where the blue lines correspond to the total variation along the horizontal edges and the red lines correspond to the total variation along the vertical edges. This splitting has already been considered before,

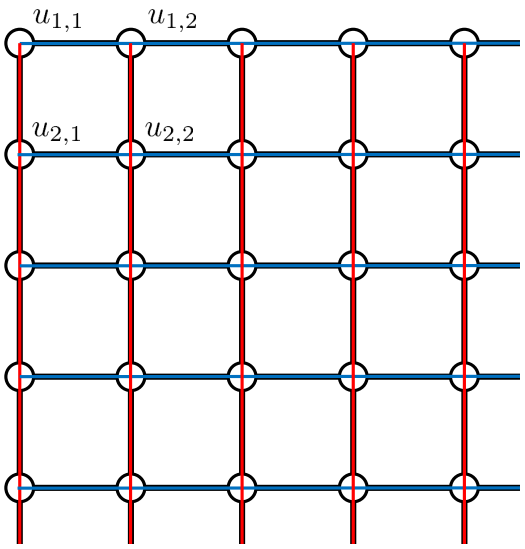


Figure 2: Chains-based splitting: The edges are decomposed into horizontal chains (blue) and vertical chains (red).

see for instance [16, 5], but to the best of our knowledge, no formal justification of the fact that it can be accelerated has been given.

We introduce auxiliary variables $u_{1,2} \in \mathbb{R}^{MN}$ and multipliers $x_{1,2} \in \mathbb{R}^{MN}$ and consider the following Lagrangian formulation of (33):

$$\min_{u_{1,2},u} \sup_{x_{1,2}} \mathrm{TV}_h(u_1) + \langle x_1, u - u_1 \rangle + \mathrm{TV}_v(u_2) + \langle x_2, u - u_2 \rangle + \frac{\lambda}{2}\|u - f\|^2.$$

Now, minimizing the Lagrangian over $u_{1,2}$ and u and denoting by $\mathrm{TV}^*_{h,v}$ the convex conjugate of $\mathrm{TV}_{h,v}$ we arrive at the dual problem

$$\max_{x_{1,2}} \mathcal{D}(x_{1,2}) = -\mathrm{TV}^*_h(x_1) - \mathrm{TV}^*_v(x_2) - \frac{1}{2\lambda}\|x_1 + x_2\|^2 + \langle x_1 + x_2, f \rangle. \tag{34}$$

The primal variable u can be recovered from the dual variables $x_{1,2}$ via

$$u = f - \frac{x_1 + x_2}{\lambda},$$

and the primal-dual gap $\mathcal{P}(u) - \mathcal{D}(x_{1,2})$ can be shown to bound $\lambda\|u - \hat{u}\|^2$, where $\hat{u}$ is the unique minimizer of (33).

Observe that (34) is exactly of the form (10) and according to Proposition 1, this problem can be accelerated. An accelerated alternating minimization takes the following form: Choose

$x_2^{-1} = x_2^0 \in \mathbb{R}^{MN}$, $t_1 = 1$, for each $k \geq 0$ compute

$$\begin{cases} t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \\ \bar{x}_2^k = x_2^k + \frac{t_k - 1}{t_{k+1}} \left(x_2^k - x_2^{k-1} \right) \\ x_1^{k+1} = \arg\min_{x_1} \mathrm{TV}_h^*(x_1) + \frac{1}{2}\|x_1 + \bar{x}_2^k - \lambda f\|^2 \\ x_2^{k+1} = \arg\min_{x_2} \mathrm{TV}_v^*(x_2) + \frac{1}{2}\|x_1^{k+1} + x_2 - \lambda f\|^2 \end{cases} . \tag{35}$$

Thanks to the Moreau identity [22, Thm 14.3], the two last lines of Algorithm (35) can be rewritten as

$$\begin{cases} x_1^{k+1} = (\lambda f - \bar{x}_2^k) - \arg\min_{x_1} \mathrm{TV}_h(x_1) + \frac{1}{2}\|x_1 - (\lambda f - \bar{x}_2^k)\|^2 \\ x_2^{k+1} = (\lambda f - x_1^{k+1}) - \arg\min_{x_2} \mathrm{TV}_v(x_2) + \frac{1}{2}\|x_2 - (\lambda f - x_1^{k+1})\|^2 \end{cases} .$$

Both partial minimization problems can be solved by solving M independent one-dimensional ROF problems on the horizontal chains and N independent one-dimensional ROF problems on the vertical chains. Efficient direct algorithms to solve one-dimensional ROF problems have been recently proposed in [16, 26, 5, 24]. The dynamic programming algorithm of [26] seems most appealing for our purpose since it guarantees a worst case complexity which is linear in the length of the chain. We will therefore make use of this algorithm.

In the experiments presented in Table 1, we compare the proposed accelerated alternating minimization (AAM) with respect to a plain alternating minimization (AM) as studied in Section 3.1. Similar to the observations made in [34], we observed that the convergence of (35) can be speeded up by restarting the extrapolation factor (by setting $t_k = 1$) of the algorithm from time to time. We experimented with different heuristics and best working heuristic turned out to restart the algorithm whenever the dual energy was increasing within the last 10 iterations. We denote this variant by (AAM-r).

In Table 2, we test a Open-MP based multi-core implementation of the (AAM-r) algorithm using a Intel Xeon CPU E5-2690 v2 @ 3.00GHz processor with 20 cores. We stop the (AAM-r) algorithm as soon as u and $x_{1,2}$ fulfill

$$\|u - \hat{u}\|_\infty \leq \|u - \hat{u}\|_2 \leq \sqrt{(\mathcal{P}(u) - \mathcal{D}(x_{1,2}))/\lambda} \leq 1/256,$$

which ensures that the maximum pixel error of u is less than $1/256$ that is the exactly accuracy of the input data. We compare the performance with a single-core implementation of the graph cut (GC) based algorithm proposed in [24, 13][2] which utilizes the max-flow algorithm of Boykov and Kolmogorov [10]. From the timings, one can observe that the proposed algorithm is already competitive to (GC) using only one core, but a multi-core implementation appears much faster. We can also observe that (AAM-r) is quite stable with respect to the value of λ.

4.2 Squares based splitting

In this section consider the a discrete approximation of the total variation on squares. Let $s = (s_1, s_2, s_3, s_4)^T$ be the nodes of a square, with s_1 being the top-left node and enumerating

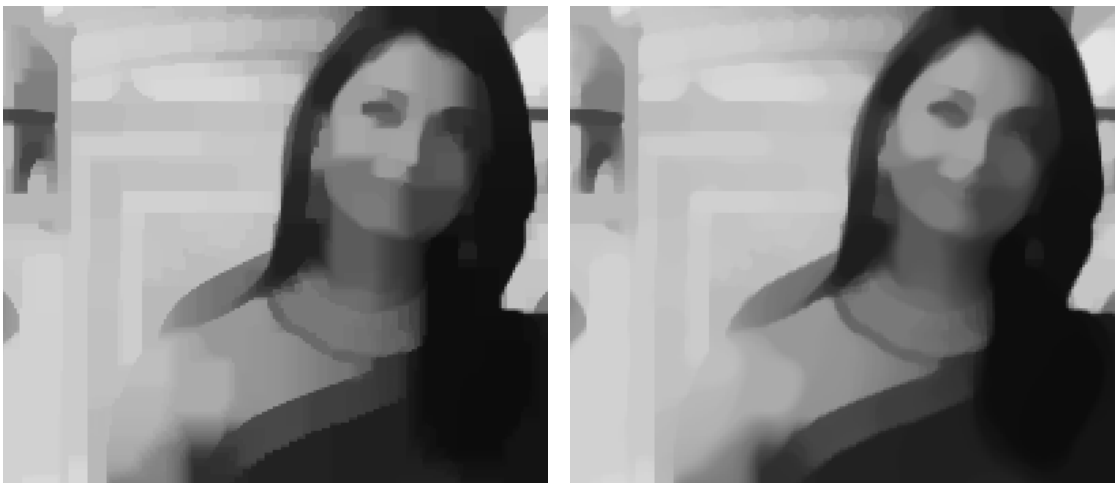
[2]The implementation has been taken from `http://www.cmap.polytechnique.fr/~antonin/software/`

Table 1: Results for chains-based splitting applied to the image shown in Figure 1. The table shows the number of iterations to reach a primal-dual gap less than `tol`.

	(AM)			(AAM)			(AAM-r)		
`tol`	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$
10^{-1}	80	90	90	30	30	30	30	30	30
10^{-3}	410	380	550	80	80	80	80	80	80
10^{-6}	1810	2220	2830	270	260	300	170	180	190
10^{-9}	3770	10000+	5640	630	790	570	220	320	250

Table 2: CPU times for the (AAM-r) algorithm using a multi-core implementation. (GC) refers to a single-core implementation of the graph cut based algorithm proposed in [24, 13].

`#cores`	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$
1	4.13	4.12	4.96
5	1.08	0.94	1.20
10	0.63	0.62	0.75
20	0.48	0.44	0.53
(GC)	3.82	6.37	19.76



(a) $p = 1$, $\lambda = 5$ (b) $p = 2$, $\lambda = 3$

Figure 3: A subview of the images shown in Figure 1 emphasizing the differences between anisotropic total variation ($p = 1$) and isotropic total variation ($p = 2$). In this example, we adapted the value of λ for a better comparison.

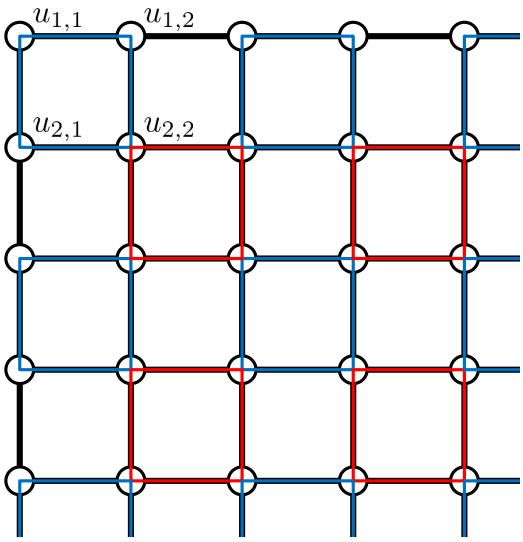


Figure 4: Squares-based splitting: The edges are decomposed into small loops (squares) where red squares have even and blue squares have odd top left indices.

the remaining nodes in clock-wise orientation. On this square we define an operator $D \in \mathbb{R}^{4\times 4}$, that computes the cyclic finite differences

$$Ds = (s_2 - s_1, s_3 - s_2, s_4 - s_3, s_1 - s_4)^T, \quad \mathrm{TV}_p(s) = \|Ds\|_p$$

and $\mathrm{TV}_p(s)$ computes the p-norm based total variation on the square s. Figure 3 shows a qualitative comparison between the anisotropic total variation ($p = 1$) and the isotropic total variation ($p = 2$).

In order to apply the definition of the total variation on squares to the whole image u, we define a linear operator $S_{i,j}$ that extracts the 4 nodes of the square from the image u with its top-left node located at (i, j), that is

$$S_{i,j}u = (u_{i,j}, u_{i,j+1}, u_{i+1,j+1}, u_{i+1,j})^T.$$

See Figure 4 for a visualization of the splitting. The idea is now to perform a splitting of the total variation into squares whose top-left nodes have even indices and squares whose top-left node have odd indices, that is

$$\mathrm{TV}_p(u) = \mathrm{TV}_e(u) + \mathrm{TV}_o(u),$$

where $\mathrm{TV}_e(u)$ corresponds to the total variation on the even squares and $\mathrm{TV}_o(u)$ corresponds to the total variation on the odd squares. They are respectively given by

$$\mathrm{TV}_e(u) = \sum_{i,j=1}^{\lfloor M/2\rfloor,\lfloor N/2\rfloor} \|DS_{2i,2j}u\|_p, \quad \mathrm{TV}_o = \sum_{i,j=1}^{\lfloor M/2\rfloor,\lfloor N/2\rfloor} \|DS_{2i-1,2j-1}u\|_p.$$

Observe that for the ease of notation, we shall skip some edges at the boundaries, which however can be easily assigned to even or odd squares. In case $p = 1$ the discretization is equivalent (up to the skipped edges at the borders) to the discretization on chains. We point out that this splitting can be extended to higher dimensions using for examples cubes with even and odd origins in 3D. Similar to the previous section we derive the dual problem as

$$\max_{x_{1,2}} \mathcal{D}(x_{1,2}) = -\operatorname{TV}_e^*(x_1) - \operatorname{TV}_o^*(x_2) - \frac{1}{2\lambda}\|x_1 + x_2\|^2 + \langle x_1 + x_2, f\rangle. \tag{36}$$

$\operatorname{TV}_{e,o}^*$ denote the conjugate functions of $\operatorname{TV}_{e,o}$. A simple computation shows that

$$\operatorname{TV}_e^*(x_1) = \sum_{i,j=1}^{\lfloor M/2\rfloor,\lfloor N/2\rfloor} \delta_K(S_{2i,2j}x_1), \quad \operatorname{TV}_o^*(x_2) = \sum_{i,j=1}^{\lfloor M/2\rfloor,\lfloor N/2\rfloor} \delta_K(S_{2i-1,2j-1}x_2),$$

where K is the convex set defined by

$$K = \{D^T\xi : \|\xi\|_q \leq 1\},$$

where $\xi \in \mathbb{R}^4$, $q = \infty$ if $p = 1$ and $q = 2$ if $p = 2$.

Observe that the conjugate functions completely decompose into independent problems on the squares. Hence, it suffices to consider the partial minimization with respect to a single square of the form

$$\min_s \delta_K(s) + \frac{1}{2}\|s - \bar{s}\|^2, \tag{37}$$

for some $\bar{s} \in \mathbb{R}^4$. Using the definition of K, this problem is equivalent to solving the constraint quadratic problem

$$\min_{\|\xi\|_q \leq 1} \frac{1}{2}\|D^T\xi - \bar{s}\|^2, \tag{38}$$

and a minimizer $\hat{s}$ of (37) can be computed from a minimizer $\hat{\xi}$ of (38) via $\hat{s} = D^T\hat{\xi}$.

4.2.1 The case $p = 1$

In case $p = 1$, all constraints on ξ are decoupled. A possibility to solve this problem is to adapt the graph cut approach [24] which in this case requires only very few computations. However, we found that it was about twice more efficient to approximately solve this problem by an alternating minimization scheme. Keeping fixed $\xi_{1,3}$, we can solve for $\xi_{2,4}$ via

$$\xi_2 = \max\left(-1, \min\left(1, \frac{\xi_1 + \xi_3 + \bar{s}_3 - \bar{s}_2}{2}\right)\right), \quad \xi_4 = \max\left(-1, \min\left(1, \frac{\xi_1 + \xi_3 + \bar{s}_1 - \bar{s}_4}{2}\right)\right).$$

Likewise keeping fixed $\xi_{2,4}$, we can globally solve for $\xi_{1,3}$ using

$$\xi_1 = \max\left(-1, \min\left(1, \frac{\xi_2 + \xi_4 + \bar{s}_2 - \bar{s}_1}{2}\right)\right), \quad \xi_3 = \max\left(-1, \min\left(1, \frac{\xi_2 + \xi_4 + \bar{s}_4 - \bar{s}_3}{2}\right)\right).$$

In a practical implementation it turns out that one iteration of this alternating minimization is enough when storing the values ξ during the iterations and performing a warm start from the previous solution.

Table 3 compares the proposed accelerated alternating minimization with a standard implementation of Beck and Teboulle's algorithm [7] (FISTA) applied to the dual problem (36). We again tested the accelerated alternating minimization algorithm (AAM) and a variant (AAM-r) that restarts the overrelaxation parameter whenever the dual energy increased within the last 100 iterations. From the results, on can see that (AAM) needs about 2-3 times less iterations and (AAM-r) needs about 3-5 times less iterations compared to (FISTA).

Table 3: Results for squares-based splitting ($p = 1$) applied to the image shown in Figure 1. The table shows the number of iterations to reach a primal-dual gap less than `tol`.

	(FISTA)			(AAM)			(AAM-r)		
`tol`	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$
10^0	800	1200	3100	300	400	1000	300	400	800
10^{-1}	1700	2900	8000	500	800	2300	500	800	1000
10^{-3}	9600	10000+	10000+	2200	4500	10000+	2200	1800	1800
10^{-6}	10000+	10000+	10000+	10000+	10000+	10000+	3900	2900	2900

4.2.2 The case $p = 2$

In case $p = 2$ we only have a single constraint $\|\xi\|_2 \leq 1$. The KKT sufficient optimality conditions of (38) are given by

$$\begin{aligned} (DD^T + \mu I)\xi - D\bar{s} &= 0 \\ \|\xi\|_2^2 - 1 &\leq 0 \\ \mu &\geq 0 \\ \mu(\|\xi\|_2^2 - 1) &= 0 \end{aligned} \tag{39}$$

where $\mu \geq 0$ is a Lagrange multiplier. Let $D = USV^T$ be a singular value decomposition of D with singular values $S = \operatorname{diag}(2, \sqrt{2}, \sqrt{2}, 0)$. Since the columns of

$$U = (u_1, u_2, u_3, u_4) = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix}$$

are the eigenvectors of DD^T we also have

$$DD^T + \mu I = U\Sigma(\mu)U^T,$$

where $\Sigma(\mu) = S^2 + \mu = \operatorname{diag}(\mu + 4, \mu + 2, \mu + 2, \mu)$. Then, the first line of (39) yields

$$\xi = U\Sigma(\mu)^+ U^T D\bar{s},$$

where $\Sigma(\mu)^+$ denotes the Moore-Penrose-Inverse of $\Sigma(\mu)$, which is well-defined also for $\mu = 0$. We deduce that

$$\|\xi\|_2^2 = \|U\Sigma(\mu)^+ U^T D\bar{s}\|_2^2 = \frac{2(t_1^2 + t_2^2)}{(\mu + 2)^2} + \frac{4t_3^2}{(\mu + 4)^2},$$

with $t_1 = \langle D^T u_2, \bar{s} \rangle$, $t_2 = \langle D^T u_3, \bar{s} \rangle$, $t_3 = \langle D^T u_1, \bar{s} \rangle$. The optimality system (39) now becomes

$$\begin{aligned} \frac{2(t_1^2 + t_2^2)}{(\mu+2)^2} + \frac{4t_3^2}{(\mu+4)^2} - 1 &\leq 0 \\ \mu &\geq 0 \\ \mu \left(\frac{2(t_1^2 + t_2^2)}{(\mu+2)^2} + \frac{4t_3^2}{(\mu+4)^2} - 1 \right) &= 0 \end{aligned}$$

We solve the reduced system for μ by a projected Newton scheme. We let $\mu^0 \geq 0$ and then, for each $n \geq 0$ we let

$$\mu^{n+1} = \max \left(0, \mu^n - \frac{\frac{2(t_1^2+t_2^2)}{(\mu+2)^2} + \frac{4t_3^2}{(\mu+4)^2} - 1}{-\frac{4(t_1^2+t_2^2)}{(\mu+2)^3} - \frac{8t_3^2}{(\mu+4)^3}} \right)$$

Once, μ is computed, ξ can be recovered from μ via

$$\xi = U \Sigma(\mu)^+ U^T D \bar{s}$$

It turns out that the above Newton scheme has a very fast convergence. If we perform a warm start from the previous solution μ during the iterations of the accelerated block descent algorithm, we observe that 6 Newton iterations are enough to reach an accuracy of 10^{-20}. In practice, the best overall performance is obtained by doing inexact optimizations using only one Newton iteration. Additionally, we can perform a simple reprojection of ξ on the constraint $\|\xi\|_p \leq 1$ before computing the dual energy to ensure feasibility of the dual problem.

Table 4 presents the results in case of the isotropic ($p = 2$) total variation on squares. In contrast to the setting $p = 1$, we observed that the restarting strategy did not significantly improve the convergence and hence we omit the results. In general, the isotropic ($p = 2$) total variation appears to be significantly more difficult to optimize compared to the anisotropic ($p = 1$) total variation. From the results it can be seen that the proposed accelerated alternating minimization (AAM) is roughly 3 times faster compared to a standard implementation of (FISTA) applied to the dual problem (36).

Table 4: Results for squares-based splitting ($p = 2$) applied to the image shown in Figure 1. The table shows the number of iterations to reach a primal-dual gap less than `tol`.

	(FISTA)			(AAM)		
`tol`	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$	$\lambda = 10$	$\lambda = 5$	$\lambda = 1$
10^0	500	700	1800	200	200	500
10^{-1}	1000	1500	3900	300	500	1200
10^{-3}	5500	8800	10000+	1500	2400	5900

Remark 2. Before closing this subsection, let us observe (cf. Section 3.3) that instead of the red-black Gauss-Seidel scheme we adopted in the two previous examples, we could also implement a standard serial Gauss-Seidel scheme, which however did not improve the results and does not allow for a parallel implementation.

4.3 Disparity estimation

In the last application we consider the problem of computing a disparity image from a pair of rectified stereo images $I^{l,r}$. We assume that $I^{l,r}$ are of size $M \times N$ and we consider K ordered disparity values $[d^1, ...d^k]$. We start from the Ishikawa formulation [25, 35] that represents the non-convex stereo problem as a minimum cut problem in a three-dimensional space.

$$\min_{\substack{u_{i,j,k+1} \leq u_{i,j,k} \\ u_{i,j,k} \in \{0,1\} \\ u_{\cdot,\cdot,1}=1 \\ u_{\cdot,\cdot,K}=0}} \mathcal{I}(u) = \mathrm{TV}_h(u) + \mathrm{TV}_v(u) + \mathrm{TV}_l(u), \tag{40}$$

where

$$\mathrm{TV}_h(u) = \sum_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N-1 \\ 1 \leq k \leq K}} w^h_{i,j} |u_{i,j+1,k} - u_{i,j,k}|, \quad \mathrm{TV}_v(u) = \sum_{\substack{1 \leq i \leq M-1 \\ 1 \leq j \leq N \\ 1 \leq k \leq K}} w^v_{i,j} |u_{i+1,j,k} - u_{i,j,k}|,$$

and

$$\mathrm{TV}_l(u) = \sum_{\substack{1 \leq i \leq M \\ 1 \leq j \leq N \\ 1 \leq k \leq K-1}} c_{i,j,k} |u_{i,j,k} - u_{i,j,k+1}|$$

The weights $w^{h,v}_{i,j}$ are edge indicator weights that are computed from the left input image I^l in order to yield improved disparity discontinuities. The weights $c_{i,j,k}$ are related to the matching cost of the left and right image for given disparity values d^k at pixel (i, j). The disparity image $d_{i,j}$ is recovered from $u_{i,j,k}$ by letting $d_{i,j} = d^k$ if $u_{i,j,k} - u_{i,j,k+1} = 1$ which can happen only for one value $k \in \{1, ..., K-1\}$.

Instead solving (40) using a max-flow algorithm as originally used in [25], we solve a 3D ROF-like problem:

$$\min_v \mathcal{P}(v) = \mathcal{I}(v) + \frac{1}{2}\|v - g\|^2,$$

where g is given for all i, j by

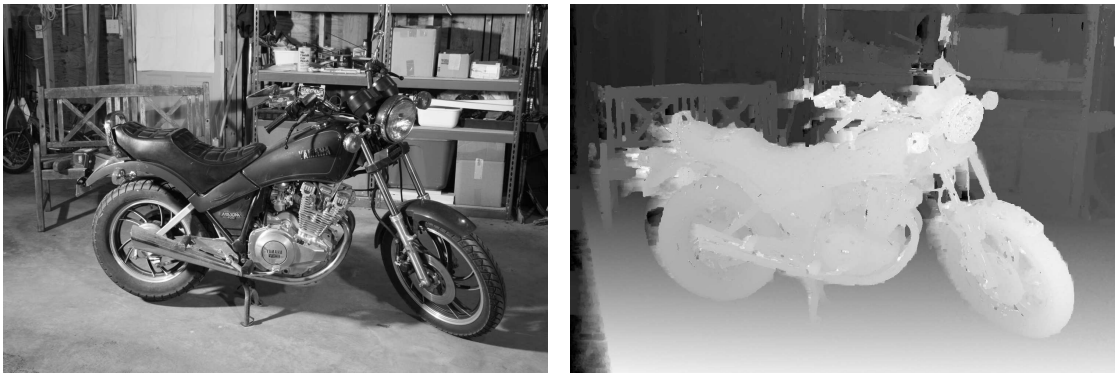
$$g_{i,j,k} = \begin{cases} \gamma & \text{if} \quad k = 0 \\ 0 & \text{if} \quad 1 < k < K \\ -\gamma & \text{if} \quad k = K \end{cases},$$

and γ is some positive constant (usually we use $\gamma = 10^3$). It can be shown that if γ is large enough, the solution $\hat{v}$ will satisfy for all i, j: $\hat{v}_{i,j,0} > 0$, and $\hat{v}_{i,j,K} < 0$. Then, in this case it can be shown [13] that

$$\hat{u}_{i,j,k} = \begin{cases} 0 & \text{if} \quad \hat{v}_{i,j,k} \geq 0 \\ 1 & \text{if} \quad \hat{v}_{i,j,k} < 0 \end{cases}$$

is a solution of (40).

We solve the 3D ROF problem by again performing a splitting into chains. Since this problem is now 3D, we need to split into three types of chains: horizontal, vertical and in the direction



(a) Left image (b) Disparity image

Figure 5: Disparity estimation: (a) shows the left input image of size 1000×1482 and (b) shows the disparity image.

of the labels. Considering a Lagrangian approach, we arrive at the dual problem

$$\max_{x_{1,2,3}} \mathcal{D}(x_{1,2,3}) = -\operatorname{TV}_h^*(x_1) - \operatorname{TV}_v^*(x_2) - \operatorname{TV}_l^*(x_3) - \frac{1}{2\lambda}\|x_1 + x_2 + x_3\|^2 + \langle x_1 + x_2 + x_3, g\rangle.$$

Since we now have three blocks, it is not clear that an accelerated alternating minimization converges (although, in fact, we observed it in practice). We should either perform plain alternating minimization, or we treat two of the variables (e.g. $x_{1,2} = (x_1, x_2)$) as one block on which we perform a partial proximal descent as investigated in Section 3. It corresponds to a particular instance of (1) with two blocks (x_1', x_2') given by $x_1' = (x_1, x_2)$ and $x_2' = x_3$ and with $A_1 x_1' = x_1 + x_2$ and $A_2 x_2' = x_3$. The functions f_1, f_2 are given by $f_1(x_1') = \operatorname{TV}_h^*(x_1) + \operatorname{TV}_v^*(x_2)$ and $f_2(x_2') = \operatorname{TV}_l^*(x_3)$. Furthermore, the step sizes are given by $\tau_1' = 1/2$ and $\tau_2' = 1$ which means that for x_1' we have to perform a descent and for x_2' we can do exact minimization.

The accelerated proximal alternating descent now takes the following form: choose $x_1^{-1} = x_1^0 \in \mathbb{R}^{MNK}$, $x_2^{-1} = x_2^0 \in \mathbb{R}^{MNK}$, set $\tau = 1/2$, and set $t_0 = 1$. For each $k \geq 0$ compute

$$\begin{cases} t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \\ \bar{x}_{1,2}^k = x_{1,2}^k + \frac{t_k - 1}{t_{k+1}}\left(x_{1,2}^k - x_{1,2}^{k-1}\right) \\ x_3^{k+1} = \arg\min_{x_3} \operatorname{TV}_l^*(x_3) + \frac{1}{2}\|\bar{x}_1^k + \bar{x}_2^k + x_3 - \lambda f\|^2 \\ x_1^{k+1} = \arg\min_{x_1} \operatorname{TV}_h^*(x_1) + \frac{1}{2\tau}\|x_1 - (\bar{x}_1^k - \tau(\bar{x}_1^k + \bar{x}_2^k + x_3^{k+1} - \lambda f))\|^2 \\ x_2^{k+1} = \arg\min_{x_2} \operatorname{TV}_v^*(x_2) + \frac{1}{2\tau}\|x_2 - (\bar{x}_2^k - \tau(\bar{x}_1^k + \bar{x}_2^k + x_3^{k+1} - \lambda f))\|^2. \end{cases} \tag{41}$$

Observe that the three proximal steps can be computed as before using an algorithm for minimizing 1D ROF problems.

We present an application to large scale disparity estimation. We use the "Motorcycle" stereo pair taken from the recently introduced high resolution stereo benchmark data set [37] at half size ($M \times N = 1000 \times 1482$). One of the two input images is shown in Figure 5 (a). We discretize the

Table 5: Results for the 3D ROF model applied to disparity estimation of the image shown in Figure 5. The table shows the iterations to reach a primal-dual gap less than `tol`. Note that due to the size of the problem, a global gap of 10^0 corresponds to a relative gap (normalized by the primal energy) of about $6.57 \cdot 10^{-10}$.

`tol`	(AM)	(AAD)
10^1	20	20
10^0	100	50
10^{-1}	390	110

disparity space in the range of $[d^1, ..., d^k] = [0, ..., 125]$ pixels. This results in $K = 126$ discrete disparity values. The weights $c_{i,j,k}$ are computed using a illumination-robust image matching cost function. For all i, j, k, we aggregate the truncated absolute differences between the image gradients of the left and right images in a 2×2 correlation window:

$$c_{i,j,k} = \frac{1}{4} \sum_{m=i-1,n=j-1}^{i,j} \min(\alpha, |(I^l_{m+1,n} - I^l_{m,n}) - (I^r_{m+1,n+d^k} - I^r_{m,n+d^k})|) + \min(\beta, |(I^l_{m,n+1} - I^l_{m,n}) - (I^r_{m,n+d^k+1} - I^r_{m,n+d^k})|).$$

The truncation values are set to $\alpha = \beta = 0.1$. The weights w^h and w^h are computed for all i, j as follows:

$$w^h_{i,j} = \lambda \cdot \begin{cases} \mu & \text{if} \quad |I^l_{i,j+1} - I^l_{i,j}| > \delta \\ 1 & \text{else} \end{cases}, \quad w^v_{i,j} = \lambda \cdot \begin{cases} \mu & \text{if} \quad |I^l_{i+1,j} - I^l_{i,j}| > \delta \\ 1 & \text{else} \end{cases},$$

where we set $\mu = \delta = 0.1$ and $\lambda = 1/600$.

Table 5 shows a comparison of the proposed accelerated alternating descent (AAD) algorithm with a standard alternating minimization (AM) algorithm which has been discussed in Section 3.1. For both algorithms we again used a multi-core implementation and ran the code on 20 cores of the same machine, mentioned above. From the table, one can see that (AAD) is much faster than (AM) especially for computing a higher accurate solution. We point out that in order to compute the disparity map the 3D ROF model does not need to be solved for a very high accuracy. Our results suggest that usually 50 iterations are enough to recover the solution of the minimum cut and hence the optimal disparity image. Note that computing the 2D disparity image amounts for computing a 3D ROF problem of size $1000 \times 1482 \times 126$, that is solving for 560196000 dual variables! One iteration of the (AAD) algorithm on the 20 core machine takes about 8.78 seconds, hence the disparity image can be computed in ~ 250 seconds. The final disparity image is shown in Figure 5 (b).

Acknowledgments

This research is partially supported by the joint ANR/FWF Project *Efficient Algorithms for Nonsmooth Optimization in Imaging (EANOI)* FWF No. I1148 / ANR-12-IS01-0003. Thomas

Pock also acknowledges support from the Austrian science fund (FWF) under the START project BIVISION, Y729. Antonin Chambolle also acknowledges support from the "Programme Gaspard Monge pour l'Optimisation" (PGMO), through the "MAORI" group.

References

[1] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2, Ser. B):5–16, 2009.

[2] H. Attouch, J. Bolte, and B. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137(1-2, Ser. A):91–129, 2013.

[3] H. Attouch, A. Cabot, P. Frankel, and J. Peypouquet. Alternating proximal algorithms for linearly constrained variational inequalities: application to domain decomposition for PDE's. *Nonlinear Anal.*, 74(18):7455–7473, 2011.

[4] A. Auslender. Asymptotic properties of the Fenchel dual functional and applications to decomposition problems. *J. Optim. Theory Appl.*, 73(3):427–449, 1992.

[5] Á. Barbero and S. Sra. Modular proximal optimization for multidimensional total-variation regularization. Technical report, arXiv:1411.0589, 2014.

[6] A. Beck. On the convergence of alternating minimization with applications to iteratively reweighted least squares and decomposition schemes. Technical Report 4154, Optimization Online, 2013.

[7] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[8] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM J. Optim.*, 23(4):2037–2060, 2013.

[9] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2, Ser. A):459–494, 2014.

[10] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, September 2004.

[11] A. Cabot and P. Frankel. Alternating proximal algorithms with asymptotically vanishing coupling. Application to domain decomposition for PDE's. *Optimization*, 61(3):307–325, 2012.

[12] P. Cannarsa and C. Sinestrari. *Semiconcave functions, Hamilton-Jacobi equations, and optimal control*. Progress in Nonlinear Differential Equations and their Applications, 58. Birkhäuser Boston, Inc., Boston, MA, 2004.

[13] A. Chambolle and J. Darbon. *Image Processing and Analysis with Graphs: Theory and Practice*, chapter A parametric maximul flow approach to discrete total variation minimization. CRC Press, 2012.

[14] A. Chambolle and C. Dossal. On the convergence of the iterates of "FISTA". submitted, September 2014.

[15] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

[16] L. Condat. A Direct Algorithm for 1-D Total Variation Denoising. *IEEE Signal Processing Letters*, 20:1054–1057, November 2013.

[17] O. Fercoq and P. Richtárik. Accelerated, parallel and proximal coordinate descent. *arXiv preprint arXiv:1312.5799*, 2013.

[18] X.L. Fu, B.S. He, X.F. Wang, and X.M. Yuan. Block-wise alternating direction method of multipliers with gaussian back substitution for multiple-block convex programming, 2014.

[19] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Oper. Res. Lett.*, 26(3):127–136, 2000.

[20] Osman Güler. New proximal point algorithms for convex minimization. *SIAM J. Optim.*, 2(4):649–664, 1992.

[21] B.S. He and X.M. Yuan. Block-wise alternating direction method of multipliers for multiple-block convex programming and beyond, 2014.

[22] P.L. Combettes H.H. Bauschke. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[23] M. Hintermüller and A. Langer. Non-overlapping domain decomposition methods for dual total variation based image denoising. *Journal of Scientific Computing*, pages 1–26, 2014.

[24] D. Hochbaum. An efficient algorithm for image segmentation, Markov random fields and related problems. *J. ACM*, 48(4):686–701 (electronic), 2001.

[25] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(10):1333–1336, 2003.

[26] N. Johnson. A dynamic programming algorithm for the fused lasso and ℓ_0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

[27] Y. Lee and A. Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. *CoRR*, abs/1305.1922, 2013.

[28] Q. Lin, Z. Lu, and L. Xiao. An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization. Technical Report MSR-TR-2014-94, Microsoft Research, July 2014.

[29] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[30] Yu. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[31] Yu. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[32] Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152, 2005.

[33] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2):341–362, 2012.

[34] B. O'Donoghue and E. Candés. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 2013.

[35] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. A convex formulation of continuous multi-label problems. In *European Conference on Computer Vision (ECCV)*, Marseille, France, October 2008.

[36] M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J. Optim.*, 23(2):1126–1153, 2013.

[37] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR 2014)*, Münster, Germany, September 2014.

[38] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008. Submitted to SIAM J. Optim.

[39] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.*, 6(3):1758–1789, 2013.