MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Parallel proximal methods for total variation minimization

Kamilov, U.

TR2016-007 March 2016

Abstract

Total variation (TV) is a widely used regularizer for stabilizing the solution of ill-posed inverse problems. In this paper, we propose a novel proximal-gradient algorithm for minimizing TV regularized least-squares cost functional. Our method replaces the standard proximal step of TV by a simpler alternative that computes several independent proximals. We prove that the proposed parallel proximal method converges to the TV solution, while requiring no sub-iterations. The results in this paper could enhance the applicability of TV for solving very large scale imaging inverse problems.

2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2016
201 Broadway, Cambridge, Massachusetts 02139

PARALLEL PROXIMAL METHODS FOR TOTAL VARIATION MINIMIZATION

Ulugbek S. Kamilov

Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA 02139, USA
Email: kamilov@merl.com

ABSTRACT

Total variation (TV) is a widely used regularizer for stabilizing the solution of ill-posed inverse problems. In this paper, we propose a novel proximal-gradient algorithm for minimizing TV regularized least-squares cost functional. Our method replaces the standard proximal step of TV by a simpler alternative that computes several independent proximals. We prove that the proposed parallel proximal method converges to the TV solution, while requiring no sub-iterations. The results in this paper could enhance the applicability of TV for solving very large scale imaging inverse problems.

***Index Terms*—** Proximal gradient method, total variation regularization, inverse problems, convex optimization

1. INTRODUCTION

The problem of estimating an unknown signal from noisy linear observations is fundamental in signal processing. The estimation task is often formulated as the linear inverse problem

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e}, \tag{1}$$

where the goal is to compute the unknown signal $\mathbf{x} \in \mathbb{R}^N$ from the noisy measurements $\mathbf{y} \in \mathbb{R}^M$. Here, the matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$ models the response of the acquisition device and the vector $\mathbf{e} \in \mathbb{R}^M$ represents the measurement noise, which is often assumed to be i.i.d. Gaussian. When the problem (1) is ill-posed, the standard approach is to rely on the regularized least-squares estimator

$$\widehat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 + \mathcal{R}(\mathbf{x}) \right\}, \tag{2}$$

where the functional $\mathcal{R}$ is a regularizer that promotes solutions with desirable properties such as transform-domain sparsity or positivity.

One of the most widely used regularizers in imaging is the total variation (TV), whose anisotropic variant can be defined as

$$\mathcal{R}(\mathbf{x}) \triangleq \lambda \|[\mathbf{D}\mathbf{x}]\|_{\ell_1} = \lambda \sum_{n=1}^{N} \sum_{d=1}^{D} |[\mathbf{D}_d \mathbf{x}]_n|, \tag{3}$$

where $\mathbf{D} : \mathbb{R}^N \to \mathbb{R}^{N \times D}$ is the discrete gradient operator, $\lambda > 0$ is a parameter controlling amount of regularization, and D is the number of dimensions in the signal. The matrix $\mathbf{D}_d$ denotes the finite difference operation along the dimension d with appropriate boundary conditions (periodization, etc.). The TV prior has been originally introduced by Rudin *et al.* [1] as a regularization approach capable of removing noise, while preserving image edges. It is often interpreted as a sparsity-promoting ℓ_1-penalty on the image gradient. TV regularization has proven to be successful in a wide range of applications in the context of sparse recovery of images from incomplete or corrupted measurements [2–8].

The minimization (2) with the TV regularization is a non-trivial optimization task. The challenging aspects are the non-smooth nature of the regularization term (3) and the massive quantity of data that typically needs to be processed. Proximal gradient methods [9] such as iterative shrinkage/thresholding algorithm (ISTA) [10–14] or alternating direction method of multipliers (ADMM) [15–17] are standard approaches to circumvent the non-smoothness of the TV regularizer.

For the optimization problem (2), ISTA can be written as

$$\mathbf{z}^t \leftarrow \mathbf{x}^{t-1} - \gamma_t \mathbf{H}^T \left(\mathbf{H}\mathbf{x}^{t-1} - \mathbf{y} \right) \tag{4a}$$

$$\mathbf{x}^t \leftarrow \operatorname{prox}_{\gamma_t \mathcal{R}}(\mathbf{z}^t), \tag{4b}$$

where $\gamma_t > 0$ is a step-size that can be determined a priori to ensure convergence [14]. Iteration (4) combines the gradient-descent step (4a) with a proximal operation (4b) defined as

$$\operatorname{prox}_{\gamma \mathcal{R}}(\mathbf{z}) \triangleq \underset{\mathbf{x} \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_{\ell_2}^2 + \gamma \mathcal{R}(\mathbf{x}) \right\}. \tag{5}$$

The proximal operator corresponds to the regularized solution of the denoising problem where $\mathbf{H}$ is an identity. Because of its simplicity, ISTA and its accelerated variants are among the methods of choice for solving practical linear inverse problems [13, 14]. Nonetheless, ISTA–based optimization of TV is complicated by the fact that the corresponding proximal operator does not admit a closed form solution. Practical implementations rely on computational solutions that require an additional nested optimization algorithm for evaluating the TV proximal [17, 18]. This typically leads to a prohibitively slow reconstruction when dealing with very large scale imaging problems such as the ones in 3D microscopy [8].

In this paper, we propose a novel approach for solving TV–based imaging problems that requires no nested iterations. This is achieved by substituting the proximal of TV with an alternative that amounts to evaluating several simpler proximal operators. One of our major contributions is the proof that the approach can achieve the true TV solution with arbitrarily high precision. We believe that the results presented in this paper are useful to practitioners working with very large scale problems that are common in 3D imaging, where the bottleneck is often in the evaluation of the TV proximal.

2. MAIN RESULTS

In this section, we present our main results. We start by introducing the proposed approach and then follow up by analyzing its convergence.

2.1. General formulation

We turn our attention to a more general optimization problem

$$\widehat{\mathbf{x}} = \operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^N} \left\{\mathcal{C}(\mathbf{x})\right\}, \tag{6}$$

where the cost functional is of the following form

$$\mathcal{C}(\mathbf{x}) = \mathcal{D}(\mathbf{x}) + \mathcal{R}(\mathbf{x}) = \mathcal{D}(\mathbf{x}) + \frac{1}{K}\sum_{k=1}^{K}\mathcal{R}_k(\mathbf{x}). \tag{7}$$

The precise connection between (7) and TV-regularized cost functional will be discussed shortly. We assume that the data-fidelity term $\mathcal{D}$ is convex and differentiable with a Lipschitz continuous gradient. This means that there exists a constant $L > 0$ such that, for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$, $\|\nabla\mathcal{D}(\mathbf{x}) - \mathcal{D}(\mathbf{z})\|_{\ell_2} \leq L\|\mathbf{x} - \mathbf{z}\|_{\ell_2}$. We also assume that each $\mathcal{R}_k$ is a continuous, convex function that is possibly nondifferentiable and that the optimal value $\mathcal{C}^*$ is finite and attained at $\mathbf{x}^*$.

We consider parallel proximal algorithms that have the following form

$$\mathbf{z}^t \leftarrow \mathbf{x}^{t-1} - \gamma_t\nabla\mathcal{D}(\mathbf{x}^{t-1}) \tag{8a}$$

$$\mathbf{x}^t \leftarrow \frac{1}{K}\sum_{k=1}^{K}\operatorname{prox}_{\gamma_t\mathcal{R}_k}(\mathbf{z}^t), \tag{8b}$$

where $\operatorname{prox}_{\gamma_t\mathcal{R}_k}$ is the proximal operator associated with $\gamma_t\mathcal{R}_k$. We are specifically interested in the case where the proximals $\operatorname{prox}_{\gamma_t\mathcal{R}_k}$ have a closed form, in which case they are preferable to the computation of the full proximal $\operatorname{prox}_{\gamma_t\mathcal{R}}$.

We now establish a connection between (7) and TV-regularized cost. Define a linear transform $\mathbf{W} : \mathbb{R}^N \rightarrow \mathbb{R}^{N\times D\times 2}$ that consists of two sub-operators: the averaging operator $\mathbf{A} : \mathbb{R}^N \rightarrow \mathbb{R}^{N\times D}$ and the discrete gradient $\mathbf{D}$ as in (3), both normlized by $1/\sqrt{2}$. The averaging operator consists of D matrices $\mathbf{A}_d$ that denote the pairwise averaging along the dimension d. Accordingly, the operator $\mathbf{W}$ is a union of scaled and shifted discrete Haar wavelet and scaling functions along each dimension [19]. Since we consider all possible shifts along each dimension the transform is redundant and can be interpreted as the union of $K = 2D$, scaled, orthogonal tranforms

$$\mathbf{W} = \begin{bmatrix}\mathbf{W}_1\\ \vdots \\ \mathbf{W}_K\end{bmatrix}. \tag{9}$$

The transform $\mathbf{W}$ and its pseudo-inverse

$$\mathbf{W}^\dagger = \frac{1}{K}[\mathbf{W}_1^T \dots \mathbf{W}_K^T] \tag{10}$$

satisfy the following two properties of Parseval frames [20]

$$\operatorname*{arg\,min}_{\mathbf{x}\in\mathbb{R}^N}\left\{\frac{1}{2}\|\mathbf{z} - \mathbf{W}\mathbf{x}\|_{\ell_2}^2\right\} = \mathbf{W}^\dagger\mathbf{z} \quad (\text{for all } \mathbf{z} \in \mathbb{R}^{KN})$$

and

$$\mathbf{W}^\dagger\mathbf{W} = \mathbf{I}. \tag{11}$$

One can thus express the TV regularizer as the following sum

$$\mathcal{R}(\mathbf{x}) = \lambda\sqrt{2}\sum_{k=1}^{K}\sum_{n\in\mathcal{H}_k}|[\mathbf{W}_k\mathbf{x}]_n|, \tag{12}$$

where $\mathcal{H}_k \subseteq [1\dots N]$ is the set of all detail coefficients of the transform $\mathbf{W}_k$. Then, the proposed parallel proximal algorithm for TV can be expressed as follows

$$\mathbf{z}^t \leftarrow \mathbf{x}^{t-1} - \gamma_t\mathbf{H}^T\left(\mathbf{H}\mathbf{x}^{t-1} - \mathbf{y}\right) \tag{13a}$$

$$\mathbf{x}^t \leftarrow \frac{1}{K}\sum_{k=1}^{K}\mathbf{W}_k^T\mathcal{T}(\mathbf{W}_k\mathbf{z}^t; \sqrt{2}K\gamma_t\lambda), \tag{13b}$$

where $\mathcal{T}$ is the component-wise shrinkage function

$$\mathcal{T}(y;\tau) \triangleq \max(|y| - \tau, 0)\frac{y}{|y|}, \tag{14}$$

which is applied only on scaled differences $\mathbf{D}\mathbf{z}^t$.

The algorithm in (13) is closely related to a technique called cycle spinning [21] that is commonly used for improving the performance of wavelet-domain denoising. In particular, when $\mathbf{H} = \mathbf{I}$ and $\gamma_t = 1$, for all $t = 1, 2, \dots$, the algorithm yields the solution

$$\widehat{\mathbf{x}} \leftarrow \mathbf{W}^\dagger\mathcal{T}(\mathbf{W}\mathbf{y}; \sqrt{2}K\lambda), \tag{15}$$

which can be interpreted as the traditional cycle spinning algorithm restricted to the Haar wavelet-transform. In the context of image denoising, the connections between TV and cycle-spinning were originally established in [22].

2.2. Theoretical convergence

The convergence results in this section assume that the gradient of $\mathcal{D}$ and subgradients of $\mathcal{R}_k$ are bounded, i.e., there exists $G > 0$ such that for all k and t, $\|\nabla\mathcal{D}(\mathbf{x}^t)\|_{\ell_2} \leq G$ and $\|\partial\mathcal{R}_k(\mathbf{x}^t)\|_{\ell_2} \leq G$. The following proposition is crucial for establishing the convergence of the parallel proximal algorithm.

Proposition 1. Consider the cost function (7) and the algorithm (8). Then, for all $t = 1, 2, \dots$, and for any $\mathbf{x} \in \mathbb{R}^N$, we have

$$\begin{aligned}&\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}) \\ &\quad\leq \frac{1}{2\gamma_t}\left(\|\mathbf{x}^{t-1}-\mathbf{x}\|_{\ell_2}^2 - \|\mathbf{x}^t - \mathbf{x}\|_{\ell_2}^2\right) + 8\gamma_t G^2.\end{aligned} \tag{16}$$

Proof: See Appendix.

Proposition 1 allows us to develop various types of convergence results. For example, if $\mathbf{x}^*$ is the optimal point and if we pick a sufficiently small step $\gamma_t \leq (\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*))/(8G^2)$, then the $\mathbf{x}^t$ will be closer to $\mathbf{x}^*$ than $\mathbf{x}^{t-1}$. This argument can be formalized into the following proposition.

Proposition 2. Assume a fixed step size $\gamma_t = \gamma > 0$. Then, we have that

$$\liminf_{t\to\infty}\left(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}^*\right) \leq 8\gamma G^2. \tag{17}$$

Proof: See Appendix.

Proposition 2 states that for a constant step-size, convergence can be established to the neighborhood of the optimimum, which can be made arbitrarily close to 0 by letting $\gamma \to 0$.

3. EXPERIMENTS

In this section, we empirically illustrate that our results hold more generally than suggested by Proposition 2. Specifically, we consider the accelerated parallel proximal algorithm based on FISTA [14]

$$\mathbf{z}^t \leftarrow \mathbf{u}^{t-1} - \gamma_t \nabla\mathcal{D}(\mathbf{u}^{t-1}) \tag{18a}$$

$$\mathbf{x}^t \leftarrow \frac{1}{K}\sum_{k=1}^{K} \mathrm{prox}_{\gamma_t \mathcal{R}_k}(\mathbf{z}^t) \tag{18b}$$

$$q_t \leftarrow (1 + \sqrt{1 + 4q_{t-1}^2})/2 \tag{18c}$$

$$\mathbf{u}^t \leftarrow \mathbf{x}^t + (q_{t-1} - 1)/q_t)(\mathbf{x}^t - \mathbf{x}^{t-1}) \tag{18d}$$

with $\mathbf{u}^0 = \mathbf{x}^0$, $q_0 = 1$, and $\gamma_t = \gamma$. Method (18) preserves the simplicity of the ISTA approach (8) but provides a significantly better rate of convergence, which enhances potential applicability of the method. We consider an estimation problem where the Shepp-Logan phantom of size 32×32 is reconstructed from $M = 512$ linear measurements with AWGN corresponding to 30 dB SNR. The measurement matrix is i.i.d. Gaussian $[\mathbf{H}]_{mn} \sim \mathcal{N}(0, 1/M)$. Figure 1 illustrates the per-iteration gap $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}^*)/\mathcal{C}^*$, where $\mathbf{x}^t$ is computed with the fast parallel proximal method (18) and $\mathcal{C}$ is the TV-regularized least-squares cost. The regularization parameter λ was manually selected for the optimal SNR performance of TV. We compare 3 different step-sizes $\gamma = 1/L$, $\gamma = 1/(4L)$, and $\gamma = 1/(16L)$, where $L = \lambda_{\text{max}}(\mathbf{H}^T\mathbf{H})$ is the Lipschitz constant. Proposition 2 suggests that the gap $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}^*)$ is proportional to the step-size and shrinks to 0 as the step-size is reduced. Such behavior is clearly observed in Figure 1, which suggests that our results potentially hold for the accelerated parallel proximal algorithm. Figure 2 compares the quality of the estimated images, for $\gamma = 1/L$ and $\gamma = 1/(16L)$, against the TV solution. We note that, even for $\gamma = 1/L$, the solution of our algorithm is very close to the true TV result, both qualitatively and quantitatively. This implies that, while requiring no nested iterations, our parallel proximal approach can potentially approximate the solution of TV with arbitrarily accurate precision at $\mathcal{O}(1/t^2)$ convergence rate of FISTA.

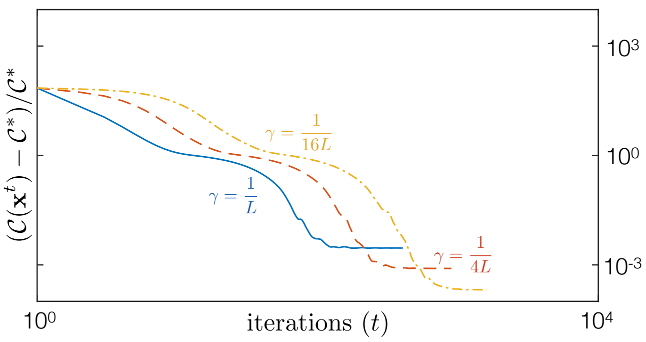

Fig. 1. Reconstruction of a *Shepp-Logan* phantom from noisy linear measurements. The relative gap $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}^*)/\mathcal{C}^*$ is plotted against the iteration number for 3 distinct step-sizes γ. The plot illustrates the convergence of the fast parallel proximal algorithm to the minimizer of the TV cost functional.

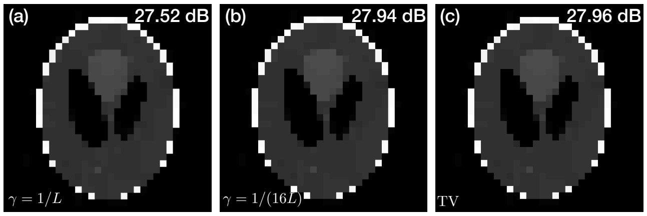

Fig. 2. Reconstructed *Shepp-Logan* images for (a) $\gamma = 1/L$; (b) $\gamma = 1/(16L)$; (c) the true TV solution. Even for $\gamma = 1/L$ the solution of the accelerated parallel proximal algorithm is visually and quantitatively close to the true TV result.

4. RELATION TO PRIOR WORK

The results in this paper are most closely related to the work on TV–based imaging by Beck and Teboulle [18]. While their approach requires additional nested optimization to compute the TV proximal, we avoid this by relying on multiple simplified proximals computed in parallel. Our proofs rely on several results from convex optimization that were used by Bertsekas [23] for analyzing a different family of algorithms called incremental proximal methods. Finally, two earlier papers with the author describe the relationship between cycle spinning and TV [22, 24], but concentrate on a fundamentally different families of optimization algorithms.

5. CONCLUSION

The parallel proximal method and its accelerated version, which were presented in this paper, are beneficial in the context of TV regularized image reconstruction, especially when the computation of the TV proximal is costly. We presented a combination of theoretical and empirical evidence demonstrating that these methods can compute the TV solution at the competitive global convergence rates without resorting to expensive sub-iterations. Future work will aim at extending the theoretical analysis presented here and by applying the methods to a larger class of imaging problems.

6. APPENDIX

We now prove the propositions in Section 2.2. The formalism used here is closely related to the analysis of incremental proximal methods that were studied by Bertsekas [23]. Related techniques were also used to analyze the convergence of recursive cycle spinning algorithm in [24].

6.1. Proof of Proposition 1

We define an intermediate quantity $\mathbf{x}_k^t \triangleq \mathrm{prox}_{\gamma_t \mathcal{R}_k}(\mathbf{z}^t)$. The optimality conditions for (8b) imply that there must exist K subgradient vectors $\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t) \in \partial\mathcal{R}_k(\mathbf{x}_k^t)$ such that

$$\mathbf{x}_k^t = \mathbf{x}^{t-1} - \gamma_t \left(\nabla\mathcal{D}(\mathbf{x}^{t-1}) + \tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t)\right). \tag{19}$$

This implies that

$$\mathbf{x}^t = \mathbf{x}^{t-1} - \gamma_t \left(\nabla\mathcal{D}(\mathbf{x}^{t-1}) + \mathbf{g}^t\right), \tag{20}$$

where

$$\mathbf{g}^t \triangleq \frac{1}{K}\sum_{k=1}^{K} \tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t).$$

Then we can write

$$\begin{aligned} \|\mathbf{x}^t - \mathbf{x}\|_{\ell_2}^2 &= \|\mathbf{x}^{t-1} - \gamma_t(\nabla\mathcal{D}(\mathbf{x}^{t-1}) + \mathbf{g}^t) - \mathbf{x}\|_{\ell_2}^2 \\ &= \|\mathbf{x}^{t-1} - \mathbf{x}\|_{\ell_2}^2 - 2\gamma_t\langle\nabla\mathcal{D}(\mathbf{x}^{t-1}) + \mathbf{g}^t, \mathbf{x}^{t-1} - \mathbf{x}\rangle \\ &\quad + \gamma_t^2\|\nabla\mathcal{D}(\mathbf{x}^{t-1}) + \mathbf{g}^t\|_{\ell_2}^2 \end{aligned} \tag{21}$$

By using the triangle inequality and noting that all the subgradients are bounded, we can bound the last term as

$$\|\nabla\mathcal{D}(\mathbf{x}^{t-1}) + \mathbf{g}^t\|_{\ell_2}^2 \leq 4G^2. \tag{22}$$

To bound the second term we proceed in two steps. We first write that

$$\begin{aligned} \langle\nabla\mathcal{D}(\mathbf{x}^{t-1}), \mathbf{x}^{t-1} - \mathbf{x}\rangle &\geq \mathcal{D}(\mathbf{x}^{t-1}) - \mathcal{D}(\mathbf{x}) \\ &\geq \mathcal{D}(\mathbf{x}^t) - \langle\nabla\mathcal{D}(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^{t-1}\rangle - \mathcal{D}(\mathbf{x}) \\ &\geq \mathcal{D}(\mathbf{x}^t) - \mathcal{D}(\mathbf{x}) - 2\gamma_t G^2, \end{aligned} \tag{23}$$

where we used the convexity of $\mathcal{D}$, the Cauchy-Schwarz inequality, and the bound on the gradients. In a similar way, we can write that

$$\begin{aligned} \langle\mathbf{g}^t, \mathbf{x}^{t-1} - \mathbf{x}\rangle &= \frac{1}{K}\sum_{k=1}^{K}\langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{x}^{t-1} - \mathbf{x}\rangle \\ &\geq \frac{1}{K}\sum_{k=1}^{K}(\mathcal{R}_k(\mathbf{x}_k^t) - \mathcal{R}_k(\mathbf{x})) - 2\gamma_t G^2 \\ &\geq \mathcal{R}(\mathbf{x}^t) - \mathcal{R}(\mathbf{x}) - 4\gamma_t G^2, \end{aligned} \tag{24}$$

where we used the convexity of $\mathcal{R}_k$s, the relationships (19) and (20), as well as bounds obtained via the Cauchy-Schwarz inequality. By plugging (22), (23), and (24) into (21) and by reorganizing the terms, we obtain the claim.

6.2. Proof of Proposition 2

By following an approach similar to Bertsekas [23], we prove the result by contradiction. Assume that (17) does not hold. Then, there must exist $\epsilon > 0$ such that

$$\liminf_{t\to\infty}(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}^*) > 8\gamma G^2 + 2\epsilon \tag{25}$$

Let $\bar{\mathbf{x}} \in \mathbb{R}^N$ be such that

$$\liminf_{t\to\infty}\mathcal{C}(\mathbf{x}^t) - 8\gamma G^2 - 2\epsilon \geq \mathcal{C}(\bar{\mathbf{x}}) \tag{26}$$

and let t_0 be large enough so that for all $t \geq t_0$, we have

$$\left|\mathcal{C}(\mathbf{x}^t) - \liminf_{t\to\infty}\mathcal{C}(\mathbf{x}^t)\right| \leq \epsilon. \tag{27}$$

By combining (26) and (27), we obtain that for all $t \geq t_0$

$$\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\bar{\mathbf{x}}) \geq 8\gamma G^2 + \epsilon. \tag{28}$$

Then from Proposition 1, for all $t \geq t_0$,

$$\begin{aligned} &\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\ell_2}^2 \\ &\quad\leq \|\mathbf{x}^{t-1} - \bar{\mathbf{x}}\|_{\ell_2}^2 - 2\gamma(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\bar{\mathbf{x}})) + 16\gamma^2 G^2 \\ &\quad\leq \|\mathbf{x}^{t-1} - \bar{\mathbf{x}}\|_{\ell_2}^2 - 2\gamma\epsilon. \end{aligned} \tag{29}$$

By iterating the inequality over t, we have for all $t \geq t_0$,

$$\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\ell_2}^2 \leq \|\mathbf{x}^{t_0} - \bar{\mathbf{x}}\|_{\ell_2} - 2(t - t_0)\gamma\epsilon, \tag{30}$$

which cannot hold for arbitrarily large t. This completes the proof.

7. REFERENCES

[1] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1–4, pp. 259–268, November 1992.

[2] M. M. Bronstein, A. M. Bronstein, M. Zibulevsky, and H. Azhari, "Reconstruction in diffraction ultrasound tomography using nonuniform FFT," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1395–1401, November 2002.

[3] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, September 2010.

[4] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[5] M. Lustig, D. L. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, December 2007.

[6] C. Louchet and L. Moisan, "Total variation denoising using posterior expectation," in *Eur. Signal Process. Conf*, Lausanne, Switzerland, August 25-29, 2008.

[7] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Adaptive total variation image deblurring: A majorization-minimization approach," *Signal Process.*, vol. 89, no. 9, pp. 1683–1693, September 2009.

[8] U. S. Kamilov, I. N. Papadopoulos, M. H. Shoreh, A. Goy, C. Vonesch, M. Unser, and D. Psaltis, "Learning approach to optical tomography," *Optica*, vol. 2, no. 6, pp. 517–522, June 2015.

[9] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2010.

[10] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, August 2003.

[11] J. Bect, L. Blanc-Feraud, G. Aubert, and A. Chambolle, "A ℓ_1-unified variational framework for image restoration," in *Proc. ECCV*, Springer, Ed., New York, 2004, vol. 3024, pp. 1–13.

[12] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, November 2004.

[13] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, December 2007.

[14] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[15] J. Eckstein and D. P. Bertsekas, "On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 1992.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[17] Z. Qin, D. Goldfarb, and S. Ma, "An alternating direction method for total variation denoising," *Optim. Method Softw.*, vol. 30, no. 3, pp. 594–615, 2015.

[18] A. Beck and M. Teboulle, "Fast gradient-based algorithm for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, November 2009.

[19] S. Mallat, *A Wavelet Tool of Signal Processing: The Sparse Way*, Academic Press, San Diego, 3rd edition, 2009.

[20] M. Unser and P. Tafti, *An Introduction to Sparse Stochastic Processes*, Cambridge Univ. Press, 2014.

[21] R. R. Coifman and D. L. Donoho, *Springer Lecture Notes in Statistics*, chapter Translation-invariant denoising, pp. 125–150, Springer-Verlag, 1995.

[22] U. S. Kamilov, E. Bostan, and M. Unser, "Wavelet shrinkage with consistent cycle spinning generalizes total variation denoising," *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 187–190, April 2012.

[23] D. P. Bertsekas, "Incremental proximal methods for large scale convex optimization," *Math. Program., Ser. B*, vol. 129, pp. 163–195, 2011.

[24] U. S. Kamilov, E. Bostan, and M. Unser, "Variational justification of cycle spinning for wavelet-based solutions of inverse problems," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1326–1330, November 2014.