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

Minimizing Isotropic Total Variation without Subiterations

Kamilov, U. S.

TR2016-109 August 2016

Abstract

Total variation (TV) is one of the most popular regularizers in the context of ill-posed image reconstruction problems. Due to its particular structure, minimization of a TV-regularized function with a fast iterative shrinkage/thresholding algorithm (FISTA) requires additional sub-iterations, which may lead to a prohibitively slow reconstruction when dealing with very large scale imaging problems. In this work, we introduce a novel variant of FISTA for isotropic TV that circumvents the need for subiterations. Specifically, our algorithm replaces the exact TV proximal with a componentwise thresholding of the image gradient in a way that ensures the convergence of the algorithm to the true TV solution with arbitrarily high precision.

International Traveling Workshop on Interactions Between Sparse Models and Technology (iTWIST)

Minimizing Isotropic Total Variation without Subiterations

Ulugbek S. Kamilov
Mitsubishi Electric Research Laboratories (MERL)
201 Broadway, Cambridge, MA 02139, USA
email: kamilov@merl.com.

Abstract— **Total variation (TV) is one of the most popular regularizers in the context of ill-posed image reconstruction problems. Due to its particular structure, minimization of a TV-regularized function with a fast iterative shrinkage/thresholding algorithm (FISTA) requires additional sub-iterations, which may lead to a prohibitively slow reconstruction when dealing with very large scale imaging problems. In this work, we introduce a novel variant of FISTA for isotropic TV that circumvents the need for subiterations. Specifically, our algorithm replaces the exact TV proximal with a componentwise thresholding of the image gradient in a way that ensures the convergence of the algorithm to the true TV solution with arbitrarily high precision.**

1 Introduction

Consider a linear inverse problem $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e}$, where the goal is to computationally reconstruct an unknown, vectorized image $\mathbf{x} \in \mathbb{R}^N$ from the noisy measurements $\mathbf{y} \in \mathbb{R}^M$ given the known matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$. The matrix $\mathbf{H}$ models the response of the acquisition device, while the unknown vector $\mathbf{e} \in \mathbb{R}^M$ represents the measurement noise. Practical inverse problems are typically ill-posed and the standard approach to reconstruct the image often relies on the regularized least-squares estimator

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathbb{R}^N} \{\mathcal{C}(\mathbf{x})\} \tag{1a}$$

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

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

One of the most popular regularizers for images is the *isotropic total-variation (TV)*

$$\mathcal{R}(\mathbf{x}) \triangleq \lambda \sum_{n=1}^{N} \|[\mathbf{D}\mathbf{x}]_n\|_{\ell_2} = \lambda \sum_{n=1}^{N} \sqrt{\sum_{d=1}^{D} ([\mathbf{D}_d \mathbf{x}]_n)^2}, \tag{2}$$

where $\lambda > 0$ is a parameter controlling the amount of the regularization, D is the number of dimensions of the signal, and $\mathbf{D} : \mathbb{R}^N \to \mathbb{R}^{N \times D}$ is the discrete gradient operator that computes finite differences along each dimension of the signal. The TV penalty has been originally introduced by Rudin *et al.* [1] as regularization approach capable of removing noise, while preserving image edges. It is often interpreted as a sparsity-promoting ℓ_1-penalty on the magnitudes of the image gradient. TV regularization has proved to be successful in a wide range of applications in the context of sparse recovery of images from incomplete or corrupted measurements [2–6].

The minimization problem (1) with the TV regularizer (2) is a non-trivial optimization task. Two challenging aspects are the non-smooth nature of the regularization term and the large size of typical vectors that need to be processed. The large-scale nature of the problem makes the direct, non-iterative reconstruction computationally unfeasible; it also restricts the iterative algorithms to the so-called first-order methods that perform reconstruction by successive applications of $\mathbf{H}$ and $\mathbf{H}^{\mathrm{T}}$. On the other hand, non-smoothness of the regularization term complicates direct application of the gradient methods. Accordingly, proximal minimization methods [7] such as *iterative shrinkage/thresholding algorithm (ISTA)* [8–10] and its accelerated variants [11, 12] are the standard first-order approaches to circumvent the non-smoothness of TV and are among the methods of choice for solving large-scale linear inverse problems.

For the general minimization problem (1), both standard and fast ISTA (often called *FISTA*) can be expressed as

$$\mathbf{x}^t \leftarrow \mathrm{prox}_{\gamma\mathcal{R}}(\mathbf{s}^{t-1} - \gamma \mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{s}^{t-1} - \mathbf{y})) \tag{3a}$$

$$\mathbf{s}^t \leftarrow \mathbf{x}^t + ((1 - q_{t-1})/q_t)(\mathbf{x}^t - \mathbf{x}^{t-1}) \tag{3b}$$

with $q_0 = 1$ and $\mathbf{x}^0 = \mathbf{s}^0 = \mathbf{x}_{\text{init}} \in \mathbb{R}^N$. Here, $\gamma > 0$ a step-size that can be set to $\gamma = 1/L$ with $L \triangleq \lambda_{\max}(\mathbf{H}^{\mathrm{T}}\mathbf{H})$ to ensure convergence and $\{q_t\}_{t \in [0,1,2,\dots]}$ are relaxation parameters [13]. For a fixed $q_t = 1$, the guaranteed global rate of convergence of (3) is $O(1/t)$, however, an appropriate choice of $\{q_t\}_{t \in [1,2,\dots]}$ leads to a faster $O(1/t^2)$ convergence, which is crucial for larger scale problems, where one tries to reduce the amount of matrix-vector products with $\mathbf{H}$ and $\mathbf{H}^{\mathrm{T}}$. Iteration (3) combines the gradient-descent step with respect to the quadratic data fidelity term with a proximal operator

$$\mathrm{prox}_{\gamma\mathcal{R}}(\mathbf{z}) \triangleq \arg\min_{\mathbf{x} \in \mathbb{R}^N} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_{\ell_2}^2 + \gamma \mathcal{R}(\mathbf{x}) \right\}. \tag{4}$$

While application of ISTA is straightforward for regularizers such as ℓ_1-penalty that admit closed form proximal operators, many other popular regularizers including TV do not have closed form proximals. This results in the need for an additional iterative algorithm for solving the corresponding minimization problem (4), which adds a significant computational overhead to the reconstruction process. For example, the original TV-FISTA by Beck and Teboulle [4] relies on an additional fast proximal-gradient algorithm for evaluating the TV proximal, which leads to sub-iterations.

In the rest of this manuscript, we describe a new variant of FISTA for solving TV regularized reconstruction problems. The algorithm builds on the traditional TV-FISTA in [4], but avoids sub-iterations by exploiting a specific approximation of the proximal as a sequence of simpler proximals. Theoretical analysis of the proposed method shows that it achieves the true TV solution with arbitrarily high precision at a global convergence rate of $O(1/t^2)$. This makes the proposed algorithm ideal for solving very large-scale reconstruction problems, where sub-iterations are undesirable.

2 Main Results

We consider the following iteration

$$\mathbf{x}^t \leftarrow \mathbf{W}^{\mathrm{T}}\mathcal{T}\left(\mathbf{W}(\mathbf{s}^{t-1} - \gamma\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{s}^{t-1} - \mathbf{y})); 2\gamma\lambda\sqrt{D}\right)$$
$$\mathbf{s}^t \leftarrow \mathbf{x}^t + ((1 - q_{t-1})/q_t)(\mathbf{x}^t - \mathbf{x}^{t-1}). \qquad (5)$$

Here, the linear transform $\mathbf{W} : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times D \times 2}$ 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 (2), both normalized by $1/(2\sqrt{D})$. Thus, $\mathbf{W}$ is a union of scaled and shifted discrete Haar wavelet and scaling functions along each dimension [14]. Since the transform can be interpreted as a union of $K = 2D$, scaled, orthogonal transforms, it satisfies $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$; however, note that $\mathbf{W}\mathbf{W}^{\mathrm{T}} \neq \mathbf{I}$ due to redundancy [15]. The nonlinear function

$$\mathcal{T}(\mathbf{z}; \tau) \triangleq \max(\|\mathbf{z}\|_{\ell_2} - \tau, 0)\frac{\mathbf{z}}{\|\mathbf{z}\|_{\ell_2}}, \qquad (6)$$

with $\mathbf{z} \in \mathbb{R}^D$, is a componenent-wise shrinkage function that is applied to each scaled difference $[\mathbf{D}\mathbf{s}]_n \in \mathbb{R}^D$, with $n \in [1, \dots, N]$. The algorithm (5) is closely related to a technique called *cycle spinning* [16] that is commonly used for impoving the performance of wavelet-domain denoising. In particular, when $\mathbf{H} = \mathbf{I}$, $\gamma = 1$, and $q_t = 1$, the algorithm yields the direct solution

$$\widehat{\mathbf{x}} \leftarrow \mathbf{W}^{\mathrm{T}}\mathcal{T}(\mathbf{W}\mathbf{y}; 2\lambda\sqrt{D}), \qquad (7)$$

which can be interpreted as the isotropic version of the traditional cycle spinning algorithm restricted to the Haar wavelet-transforms. Additionally, (5) is an extension of the parallel proximal algorithm in [17] to the isotropic TV implemented with FISTA.

To establish convergence of the method (5), one remarks the following equivalence (see also the relevant discussion in [18])

$$\min_{\mathbf{x} \in \mathbb{R}^N}\left\{\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 + \lambda\sum_{n=1}^{N}\|[\mathbf{D}\mathbf{x}]_n\|_{\ell_2}\right\}$$
$$= \min_{\mathbf{u} \in \mathbb{R}^{KN}}\left\{\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{u}\|_{\ell_2}^2 + \mathcal{R}_1(\mathbf{u}) + \mathcal{R}_2(\mathbf{u})\right\}, \qquad (8)$$

where the regularizers are defined as

$$\mathcal{R}_1(\mathbf{u}) \triangleq 2\lambda\sqrt{D}\sum_{n=1}^{N}\|[\mathbf{u}^{\text{dif}}]_n\|_{\ell_2} \text{ and } \mathcal{R}_2(\mathbf{u}) \triangleq \mathbb{1}_{\mathcal{U}}(\mathbf{u}). \qquad (9)$$

Here, $\mathbf{u}^{\text{dif}}$ denotes difference coefficients of $\mathbf{u} = \mathbf{W}\mathbf{x}$, and $\mathbb{1}_{\mathcal{U}}$ is the indicator function for the set

$$\mathcal{U} \triangleq \{\mathbf{u} \in \mathbb{R}^{KN} : \mathbf{u} = \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{u}\}. \qquad (10)$$

Thus, the algorithm (5) can be interpreted as a simple *incremental proximal-gradient algorithm* [19] that approximates the proximal of $\mathcal{R}$ in (2) with the successive evaluation of two proximals of $\mathcal{R}_1$ and $\mathcal{R}_2$ in (9). Then, by assuming that the gradient of the data-term and subgradients of $\mathcal{R}_1$ and $\mathcal{R}_2$ are bounded for every iterate, one can establish the following proposition.

Proposition 1. Denote with $\mathbf{x}^*$ the solution of (1) with the TV regularizer (2). Then, for an appropriate choice of $\{q_t\}_t$, the iterates generated by the proposed method in (5) satisfies

$$\left(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*)\right) \leq \frac{2}{\gamma(t+1)^2}\|\mathbf{x}^0 - \mathbf{x}^*\|_{\ell_2}^2 + \gamma G^2, \qquad (11)$$

where $G > 0$ is a fixed constant proportional to the bound on the gradients.

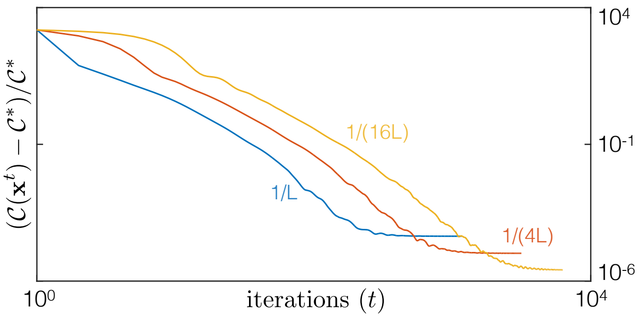

Figure 1: Recovery of the *Shepp-Logan* phantom from blurry and noisy measurements. Top: We plot the relative gap $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*))/\mathcal{C}(\mathbf{x}^*)$ against the iteration number for 3 distinct step-sizes γ. Bottom: Visual illustration of the final results. The figure illustrates the convergence of the proposed method to the minimizer of the TV cost functional. Even for $\gamma = 1/L$ the solution of the proposed method is visually and quantitatively close to the true TV result.

Note that appropriate choice of $\{q_t\}_{t \in [1,2,\dots]}$, simply refers to the choice of relaxation parameters used in the standard FISTA. The proof of the proposition can be established by extending the original proof of FISTA in [12] to sums of proximals as was done, for example, in [17] and [20].

Proposition 1 states that for a constant step-size, convergence can be established to a neighborhood of the optimum, which can be made arbitrarily close to 0 by letting $\gamma \rightarrow 0$. Additionally, the global convergence rate of the method equals that of the original TV-FISTA.

In Fig. 1, we illustrate the results of a simple image deblurring problem, where a 3 × 3 Gaussian blur of variance 2 was applied to the *Shepp-Logan* phantom. The blurry image was further contaminated with additive white Gaussian noise (AWGN) of 35 dB SNR. We plot the per-iteration gap $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*))/\mathcal{C}(\mathbf{x}^*)$, where $\mathbf{x}^t$ is computed with the proposed algorithm and $\mathbf{x}^*$ is the actual TV result. 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 $L = \lambda_{\text{max}}(\mathbf{H}^{\mathrm{T}}\mathbf{H})$ is the Lipschitz constant. The figure illustrates that the theoretical results in Proposition 1 hold in practice, and that even for $\gamma = 1/L$ the solution of the proposed method is very close to the true TV result, both qualitatively and quantitatively. For this experiment, it takes about 4 times less time (in seconds) to compute the TV solution with the proposed method compared to the standard TV-FISTA.

To conclude, we proposed to use the method summarized in eq. (5) as a fast alternative to the original TV-FISTA. The time gains of the method come from the fact that it has $O(1/t^2)$ global rate of convergence and that it uses a closed form proximal instead of solving an inner minimization problem.

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. 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.

[4] 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.

[5] D. Liu, U. S. Kamilov, and P. T. Boufounos, "Sparsity-driven distributed array imaging," in *Proc. 6th Int. Workshop on Computational Advances in Multi-Sensor Adaptive Process. (CAMSAP 2015)*, Cancun, Mexico, December 13-16, 2015, pp. 441–444.

[6] U. S. Kamilov, I. N. Papadopoulos, M. H. Shoreh, A. Goy, C. Vonesch, M. Unser, and D. Psaltis, "Optical tomographic image reconstruction based on beam propagation and sparse regularization," *IEEE Trans. Comp. Imag.*, vol. 2, no. 1, pp. 59–70,, March 2016.

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

[8] 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.

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

[10] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, November 2004.

[11] 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.

[12] 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.

[13] ——, *Convex Optimization in Signal Processing and Communications*. Cambridge, 2009, ch. Gradient-Based Algorithms with Applications to Signal Recovery Problems, pp. 42–88.

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

[15] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, pp. 947–968, 2007.

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

[17] U. S. Kamilov, "Parallel proximal methods for total variation minimization," in *IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP 2016)*, Shanghai, China, March 19-25, 2015, pp. 4697–4701.

[18] 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.

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

[20] H. H. Bauschke, R. Goebel, Y. Lucet, and X. Wang, "The proximal average: Basic theory," *SIAM J. Optim.*, vol. 19, no. 2, pp. 766–785, 2008.