A Parallel Proximal Algorithm for Anisotropic Total Variation Minimization

Kamilov, U.

TR2017-002 February 2017

Abstract

Total variation (TV) is a one of the most popular regularizers for stabilizing the solution of ill-posed inverse problems. This paper proposes a novel proximal-gradient algorithm for minimizing TV regularized least-squares cost functionals. Unlike traditional methods that require nested iterations for computing the proximal step of TV, our algorithm approximates the latter with several simple proximals that have closed form solutions. We theoretically prove that the proposed parallel proximal method achieves the TV solution with arbitrarily high precision at a global rate of converge that is equivalent to the fast proximal gradient methods. The results in this paper have the potential to enhance the applicability of TV for solving very large scale imaging inverse problems.

IEEE Transactions on Image Processing

A Parallel Proximal Algorithm for Anisotropic Total Variation Minimization

Ulugbek S. Kamilov, *Member, IEEE*

***Abstract*—Total variation (TV) is a one of the most popular regularizers for stabilizing the solution of ill-posed inverse problems. This paper proposes a novel proximal-gradient algorithm for minimizing TV regularized least-squares cost functionals. Unlike traditional methods that require nested iterations for computing the proximal step of TV, our algorithm approximates the latter with several simple proximals that have closed form solutions. We theoretically prove that the proposed parallel proximal method achieves the TV solution with arbitrarily high precision at a global rate of converge that is equivalent to the fast proximal-gradient methods. The results in this paper have the potential to 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**

I. 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 that consists in the minimization of a cost functional. The latter typically includes a quadratic data-fidelity term, as well as a regularizer that mitigates the ill-posedness of the problem by promoting solutions with desirable properties such as transform-domain sparsity or positivity.

One of the most widely used regularizers in the context of image reconstruction is total variation (TV). TV was originally introduced by Rudin *et al.* [1] as a regularization approach capable of reducing noise, while preserving image edges. It is often interpreted as a sparsity-promoting ℓ_1-penalty on the image gradient and 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 of TV regularized cost functionals is a nontrivial optimization task. The challenging aspects are the non-smooth nature of the regularization term and the large amount of data that needs to be processed in a typical application. Proximal gradient methods [9] such as iterative shrinkage/thresholding algorithm (ISTA) [10]–[12] and its accelerated variants [13], [14] are standard approaches to circumvent the non-smoothness of the TV regularizer and are among the methods of choice for solving practical linear inverse problems.

Nonetheless, ISTA–based optimization of TV is complicated by the fact that the corresponding proximal operator does not have a closed form solution. Practical implementations rely on computational solutions that require an additional nested optimization algorithm for evaluating the TV proximal [15], [16]. This typically leads to a prohibitively slow reconstruction when dealing with very large scale imaging problems such as the ones in 3D computational imaging.

In this paper, we propose a novel approach for solving TV–based imaging problems that requires no nested iterations. We consider anisotropic variant of TV and eliminate sub-iterations by approximating the exact proximal with an alternative that evaluates several simpler proximal operators that have closed form solutions. Conceptually, our method builds upon two distinct lines of prior research on inexact proximal-gradient algorithms [17]–[20] and cycle spinning [10], [21]–[23]. We believe that the results presented in this paper are useful to practitioners working with very large scale problems where the bottleneck is in the evaluation of the TV proximal.

Two key contributions of this paper are summarized as follows

- New parallel proximal-gradient method for solving anisotropic TV regularized reconstruction problems. The algorithm builds upon fast iterative shrinkage/thresholding algorithm (FISTA) [15], but avoids sub-iterations by exploiting a specific decomposition of TV as an average of several simple regularizers.
- Theoretical analysis of the method proving that it achieves the TV solution with arbitrarily high precision at a global convergence rate of $\mathcal{O}(1/t^2)$, where t denotes the iteration number. This makes the proposed algorithm ideal for solving very large-scale image reconstruction problems, where nested optimization is undesirable. In addition, we experimentally illustrate possible computational gains due to our approach on the problems of image deconvolution and super-resolution.

A. Related Work

The results in this paper are most closely related to the work on TV–based image reconstruction by Beck and Teboulle [15]. Their approach for solving TV requires an additional nested FISTA, implemented in the dual form, for evaluating the proximal. Our aim is to avoid sub-iterations by replacing the exact TV proximal with a specific approximation that can still guarantee convergence to the TV solution. While Beck and Teboulle's approach considers both isotropic and anisotropic variants of the TV reguralization, we restrict our attention to anisotropic TV.

In another related work, Condat [24] proposed a direct algorithm for 1D TV proximal, which can be used to accelerate the

U. S. Kamilov is with Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, 8th floor, Cambridge, MA 02139, USA (e-mail: kamilov@merl.com)

resolution of 1D TV-regularized inverse problems. Approach taken in this paper is fundamentally different, where instead of finding an exact computational solution for evaluating the proximal, we find a suitable approximation. This, however, allows our method to generalize to inverse problems of arbitrary number of dimensions.

From the convex optimization perspective, our work is related to inexact proximal-gradient algorithms that were extensively studied for various applications. For example, in the context of online learning, Zinkevich [25] has proposed an incremental projected-gradient algorithm that minimizes a smooth cost functional by evaluating its partial gradients. He has proved that, with a proper adaptation of the step size, the algorithm reaches the minimizer at a global convergence rate of $\mathcal{O}(1/\sqrt{t})$. The algorithm and its analysis were extended by Duchi and Singer [26] for optimizing cost functionals containing non-smooth regularizers. Bertsekas [18] further generalized those results to include algorithms that combine partial gradient, subgradient, and proximal iterations. D'Aspremont [17] showed that optimal $\mathcal{O}(1/t^2)$ complexity of Nesterov's algorithm [27] is preserved, when the gradient is computed only up to a small, uniformly bounded error. More recently, Schmidt [19] and Devolder *et al.* [20] have investigated the convergence rates of proximal-gradient algorithms when proximals are approximated iteratively.

The results in this paper are also related to a technique called cycle spinning that is commonly used for improving the performance of wavelet-domain regularization [28]. The concept was first introduced by Coifman and Donoho [21] for wavelet-domain denoising. The recursive approach to cycle spinning was studied by Fletcher *et al.* [29]. Cycle spinning was then applied to more-general linear inverse problems by Figueiredo and Nowak [10]. Currently, it is used in the majority of waveled-based reconstution algorithms to obtain higher-quality solutions with less-blocky artifacts [30]–[33]. Finally, two earlier papers with the author established the relationship between cycle spinning and TV [22], [23], but concentrated on different types of optimization algorithms. Those prior works have inspired the preliminary version of this paper [34] that concentrated on the standard ISTA without acceleration. This paper significantly extends [34] by providing extensive theoretical and empirical justifications of fast parallel proximal algorithm for TV minimization.

II. Background

A. Problem Formulation

We consider a 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 formulate the estimation as the following minimization problem

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \{\mathcal{C}(\mathbf{x})\} \tag{2}$$

$$= \arg\min_{\mathbf{x}\in\mathbb{R}^N} \{\mathcal{D}(\mathbf{x}) + \mathcal{R}(\mathbf{x})\} \tag{3}$$

where

$$\mathcal{D}(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 \tag{4}$$

is the quadratic data fidelity term. Two common variants of TV are the anisotropic TV regularizer

$$\mathcal{R}(\mathbf{x}) \triangleq \lambda\|\mathbf{D}\mathbf{x}\|_{\ell_1} = \lambda\sum_{n=1}^{N}\sum_{d=1}^{D}|[\mathbf{D}_d\mathbf{x}]_n| \tag{5}$$

and isotropic TV regularizer

$$\mathcal{R}(\mathbf{x}) \triangleq \lambda\sum_{n=1}^{N}\|[\mathbf{D}\mathbf{x}]_n\|_{\ell_2} = \lambda\sum_{n=1}^{N}\sqrt{\sum_{d=1}^{D}([\mathbf{D}_d\mathbf{x}]_n)^2}. \tag{6}$$

Here, $\mathbf{D} : \mathbb{R}^N \rightarrow \mathbb{R}^{N\times D}$ 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, Neumann boundary conditions, etc.).

B. Fast Iterative Shrinkage/Thresholding Algorithm

One popular approach for solving (2) is ISTA

$$\mathbf{x}^t \leftarrow \mathrm{prox}_{\gamma_t\mathcal{R}}(\mathbf{x}^{t-1} - \gamma_t\nabla\mathcal{D}(\mathbf{x}^{t-1})), \tag{7}$$

where the gradient of the quadratic term is given by

$$\nabla\mathcal{D}(\mathbf{x}) = \mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{x} - \mathbf{y}) \tag{8}$$

and $\gamma_t > 0$ is a step-size that can be set to $\gamma_t = 1/L$ with $L \triangleq \lambda_{\max}(\mathbf{H}^{\mathrm{T}}\mathbf{H})$ to ensure convergence [14]. Iteration (7) combines the gradient-descent step with a proximal operation defined as

$$\mathrm{prox}_{\gamma\mathcal{R}}(\mathbf{z}) \triangleq \arg\min_{\mathbf{x}\in\mathbb{R}^N}\left\{\frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_{\ell_2}^2 + \gamma\mathcal{R}(\mathbf{x})\right\}. \tag{9}$$

The proximal operator (9) corresponds to the regularized solution of the denoising problem where $\mathbf{H}$ is an identity.

Although elegant and simple, it is well known that ISTA only achieves a suboptimal convergence rate of $\mathcal{O}(1/t)$. Its accelerated version FISTA can be described as follows

$$\mathbf{x}^t \leftarrow \mathrm{prox}_{\gamma_t\mathcal{R}}(\mathbf{u}^{t-1} - \gamma_t\nabla\mathcal{D}(\mathbf{u}^{t-1})) \tag{10a}$$

$$q_t \leftarrow \left(1 + \sqrt{1 + 4q_{t-1}^2}\right)/2 \tag{10b}$$

$$\mathbf{u}^t \leftarrow \mathbf{x}^t + (q_{t-1} - 1)/q_t)(\mathbf{x}^t - \mathbf{x}^{t-1}) \tag{10c}$$

with $\mathbf{u}^0 = \mathbf{x}^0$ and $q_0 = 1$. Method (10) preserves the simplicity of ISTA (7), but provides a significantly better rate of convergence as summarized in the following theorem from [14].

Theorem 1. Assume a fixed step size $\gamma_t = \gamma \in (0, 1/L]$ and let $\{\mathbf{x}^t\}_{t=1,2,\ldots}$ be the sequence of estimates generated by FISTA. Then for any $t \geq 1$, we have that

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

where $\mathbf{x}^*$ is a minimizer of $\mathcal{C}$.

The change in convergence rate from $\mathcal{O}(1/t)$ to $\mathcal{O}(1/t^2)$ becomes crucial when solving very large scale inverse problems, where one tries to reduce the amount of matrix-vector products with $\mathbf{H}$ and $\mathbf{H}^{\mathrm{T}}$.

Application of ISTA and FISTA is straightforward for regularizers such as ℓ_1-penalty that admit closed form proximal operators (9). However, many other popular regularizers including TV do not have closed form proximals and require an additional iterative algorithm for solving (9). This adds significant computational overhead to the estimation process, which we shall eliminate for the anisotropic TV in the next section.

III. Proposed Approach

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

A. General formulation

We turn our attention to a more general optimization problem

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\arg\min}\left\{\mathcal{C}(\mathbf{x})\right\}, \tag{12}$$

where the cost functional is of the following form

$$\mathcal{C}(\mathbf{x}) = \mathcal{D}(\mathbf{x}) + \mathcal{R}(\mathbf{x}) = \mathcal{D}(\mathbf{x}) + \frac{1}{K}\sum_{k=1}^{K}\mathcal{R}_k(\mathbf{x}). \tag{13}$$

The precise connection between (13) 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 $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$, $\|\nabla\mathcal{D}(\mathbf{x}) - \nabla\mathcal{D}(\mathbf{z})\|_{\ell_2} \leq L\|\mathbf{x} - \mathbf{z}\|_{\ell_2}$. 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 $\mathbf{x}^*$ (which is not necessarily unique).

We consider fast parallel proximal algorithms that have the following form

$$\mathbf{x}^t \leftarrow \frac{1}{K}\sum_{k=1}^{K}\mathrm{prox}_{\gamma_t\mathcal{R}_k}(\mathbf{u}^{t-1} - \gamma_t\nabla\mathcal{D}(\mathbf{u}^{t-1})) \tag{14a}$$

$$q_t \leftarrow \left(1 + \sqrt{1 + 4q_{t-1}^2}\right)/2 \tag{14b}$$

$$\mathbf{u}^t \leftarrow \mathbf{x}^t + (q_{t-1} - 1)/q_t)(\mathbf{x}^t - \mathbf{x}^{t-1}), \tag{14c}$$

with $\mathbf{u}^0 = \mathbf{x}^0$ and $q_0 = 1$. Here, $\mathrm{prox}_{\gamma_t\mathcal{R}_k}$ is the proximal operator associated with $\gamma_t\mathcal{R}_k$. We are specifically interested in the case where the proximals $\mathrm{prox}_{\gamma_t\mathcal{R}_k}$ have closed forms, in which case they are preferable to the computation of the full proximal $\mathrm{prox}_{\gamma_t\mathcal{R}}$.

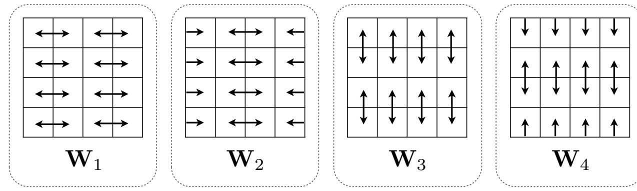

Fig. 1. Decomposition of a 4×4 image into 4 orthogonal wavelets $\{\mathbf{W}_k\}_{k\in[1,\ldots,4]}$ with periodic boundary conditions. Arrows indicate pixels used for computing averages and differences for the transforms.

We now establish a connection between (13) and TV-regularized cost. Define a linear transform $\mathbf{W} : \mathbb{R}^N \to \mathbb{R}^{N\times D\times 2}$ that consists of two sub-operators: the averaging operator $\mathbf{A} : \mathbb{R}^N \to \mathbb{R}^{N\times D}$ and the discrete gradient $\mathbf{D}$ as in (5), both normalized 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 [28]. 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

$$\mathbf{W} = \begin{bmatrix}\mathbf{W}_1\\ \vdots \\ \mathbf{W}_K\end{bmatrix}. \tag{15}$$

Figure 1 illustrates the grouping of differences and averages into 4 wavelets for a $2D$ image. The transform $\mathbf{W}$ and its pseudo-inverse

$$\mathbf{W}^\dagger = \frac{1}{K}[\mathbf{W}_1^{\mathrm{T}} \ldots \mathbf{W}_K^{\mathrm{T}}] \tag{16}$$

satisfy the following two properties of Parseval frames [35]

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

and

$$\mathbf{W}^\dagger\mathbf{W} = \mathbf{I}. \tag{17}$$

One can thus express the anisotropic TV regularizer as the following sum

$$\mathcal{R}(\mathbf{x}) = \lambda\sqrt{2}\sum_{k=1}^{K}\sum_{n\in\mathcal{H}_k}|[\mathbf{W}_k\mathbf{x}]_n|, \tag{18}$$

where $\mathcal{H}_k \subseteq [1 \ldots N]$ 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

$$\mathbf{z}^t \leftarrow \mathbf{x}^{t-1} - \gamma_t\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{x}^{t-1} - \mathbf{y}\right) \tag{19a}$$

$$\mathbf{x}^t \leftarrow \frac{1}{K}\sum_{k=1}^{K}\mathbf{W}_k^{\mathrm{T}}\mathcal{T}(\mathbf{W}_k\mathbf{z}^t; \sqrt{2}K\gamma_t\lambda), \tag{19b}$$

and fast parallel proximal algorithm can be expressed as

$$\mathbf{z}^t \leftarrow \mathbf{u}^{t-1} - \gamma_t \mathbf{H}^{\mathrm{T}} \left(\mathbf{H}\mathbf{u}^{t-1} - \mathbf{y}\right) \tag{20a}$$

$$\mathbf{x}^t \leftarrow \frac{1}{K} \sum_{k=1}^{K} \mathbf{W}_k^{\mathrm{T}} \mathcal{T}(\mathbf{W}_k \mathbf{z}^t; \sqrt{2} K \gamma_t \lambda) \tag{20b}$$

$$q_t \leftarrow (1 + \sqrt{1 + 4q_{t-1}^2})/2 \tag{20c}$$

$$\mathbf{u}^t \leftarrow \mathbf{x}^t + (q_{t-1} - 1)/q_t)(\mathbf{x}^t - \mathbf{x}^{t-1}), \tag{20d}$$

with $\mathbf{u}^0 = \mathbf{x}^0$ and $q_0 = 1$. Here, $\mathcal{T}$ is the component-wise shrinkage function

$$\mathcal{T}(y; \tau) \triangleq \max(|y| - \tau, 0) \frac{y}{|y|}, \tag{21}$$

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

The algorithm in (19) 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

$$\widehat{\mathbf{x}} \leftarrow \mathbf{W}^{\dagger} \mathcal{T}(\mathbf{W}\mathbf{y}; \sqrt{2} K \lambda), \tag{22}$$

which can be interpreted as the traditional cycle spinning algorithm restricted to the the Haar wavelet-transform.

B. 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 that we prove in the appendix establishes the convergence of the fast parallel proximal algorithm.

Proposition 1. Assume a fixed step size $\gamma_t = \gamma \in (0, 1/L]$. Then, we have that

$$\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*) \leq \frac{2}{\gamma(t+1)^2} \|\mathbf{x}^0 - \mathbf{x}^*\|_{\ell_2}^2 + 4\gamma G^2. \tag{23}$$

Proof: See Appendix.

Proposition 1 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 $\gamma \to 0$. Additionally, the global convergence rate of fast parallel proximal algorithm matches that of FISTA. Note that the result here extends the preliminary work [34] that established the convergence of the standard parallel proximal algorithm (19).

IV. Numerical Examples

The main purpose of this section is to empirically demonstrate the convergence of our fast parallel proximal algorithm (FPPA) and validate our theoretical contribution in Proposition 1. We additionally present some comparisons of TV against some other state-of-the-art methods on the problem of single image super-resolution. All the simulations were performed with MATLAB on an Apple iMac with a 4 GHz Intel Core i7 processor and 32 GBs of memory.

Fig. 2. Four examples from the standard image dataset `Set14`. From left to right: *Man*, *Baboon*, *Barbara*, and *Coastguard*.

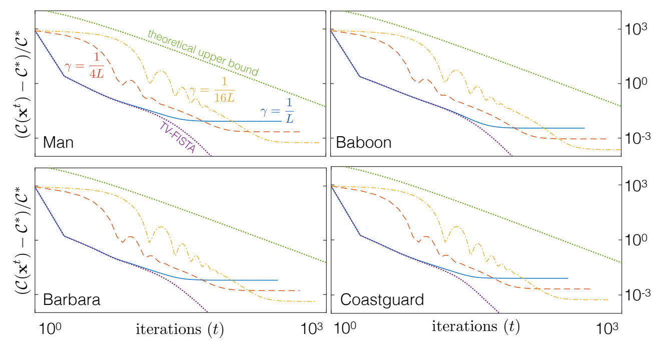

Fig. 3. Recovery of test images from blurry and noisy measurements. The relative cost accuracy $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*))/\mathcal{C}(\mathbf{x}^*)$ is plotted against the iteration number for 3 distinct step-sizes γ. The top dotted line corresponds to the theoretical upper bound (11), while the bottom one plots the evolution of the relative cost accuracy of the exact TV-FISTA. This plot illustrates the accuracy of FPPA relative to the minimizer of the TV cost functional.

A. Empirical Validation of Proposition 1

To empirically validate the convergence of the algorithm, we consider an image deblurring problem where the blur is a 5×5 Gaussian of variance 2 and where the blurry image is contaminated with an additive white Gaussian noise (AWGN) of 30 dB SNR. We evaluate the performance on a standard image dataset `Set14`, used in the previous works [36]–[38]. Following these works, only the luminance component of color images was considered. Some examples from the dataset are illustrated in Figure 2.

The simulation results on all 14 images are summarized in Table I. There we compare FPPA against the exact TV-FISTA [15], which computes the TV proximal iteratively in the dual domain. For FPPA, we consider 3 different step-sizes $\gamma = 1/L$, $\gamma = 1/(4L)$, and $\gamma = 1/(16L)$, where $L = \lambda_{\max}(\mathbf{H}^{\mathrm{T}}\mathbf{H})$ is the Lipschitz constant, and report three quantities: (a) the relative cost accuracy $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*))/\mathcal{C}(\mathbf{x}^*)$, (b) the relative peak signal-to-noise ratio (PSNR) in dB with respect to the TV solution $\mathbf{x}^*$, and (c) the speedup factor. The images $\mathbf{x}^t$ and $\mathbf{x}^*$ are computed with FFPA and the exact TV-FISTA, respectively, and $\mathcal{C}$ is the TV-regularized least-squares cost. The regularization parameter λ was manually selected for the optimal PSNR performance of TV. To ensure the convergence, we deliberately select the same strict stopping rules for all the algorithms; they are run for a maximum of $t_{\max} = 10^4$ iterations with an additional stopping criterion based on measuring the relative change of the solution in two successive iterations

$$\frac{\|\mathbf{x}^t - \mathbf{x}^{t-1}\|_{\ell_2}}{\|\mathbf{x}^{t-1}\|_{\ell_2}} \leq 10^{-5}. \tag{24}$$

The maximal number of inner iterations for the proximal of

TABLE I
THE RELATIVE COST, PSNR WITH RESPECT TO THE TV SOLUTION, AND THE SPEEDUP FACTOR FOR THE IMAGES FROM Set14.

Set14	$\gamma = 1/L$			$\gamma = 1/(4L)$			$\gamma = 1/(16L)$		
Images	Cost Accuracy	PSNR rel. TV	Time	Cost Accuracy	PSNR rel. TV	Time	Cost Accuracy	PSNR rel. TV	Time
Baboon	0.0035	48.40	×16	0.0009	58.45	×10	0.0002	65.52	×7
Barbara	0.0062	48.01	×19	0.0016	57.24	×11	0.0004	65.89	×7
Bridge	0.0089	45.66	×20	0.0023	55.14	×12	0.0006	64.89	×8
Coastguard	0.0081	46.81	×14	0.0021	55.95	×8	0.0005	65.23	×5
Comic	0.0087	43.32	×13	0.0023	52.65	×8	0.0006	62.13	×5
Face	0.0290	43.84	×13	0.0083	51.70	×7	0.0022	61.21	×5
Flowers	0.0097	47.13	×12	0.0026	56.32	×8	0.0007	65.42	×5
Foreman	0.0130	43.03	×15	0.0037	51.23	×9	0.0010	60.38	×5
Lenna	0.0260	41.06	×22	0.0080	48.67	×12	0.0021	57.70	×8
Man	0.0084	47.61	×17	0.0022	56.88	×10	0.0006	66.22	×7
Monarch	0.0200	43.73	×19	0.0058	51.85	×12	0.0015	60.96	×8
Pepper	0.0130	44.31	×19	0.0038	52.66	×12	0.0010	61.84	×8
Ppt3	0.0160	41.69	×16	0.0047	49.50	×9	0.0012	58.03	×6
Zebra	0.0160	41.49	×13	0.0044	50.29	×8	0.0011	59.95	×5
Average	**0.0133**	**44.72**	**×16**	**0.0038**	**53.47**	**×10**	**0.0009**	**62.53**	**×6**

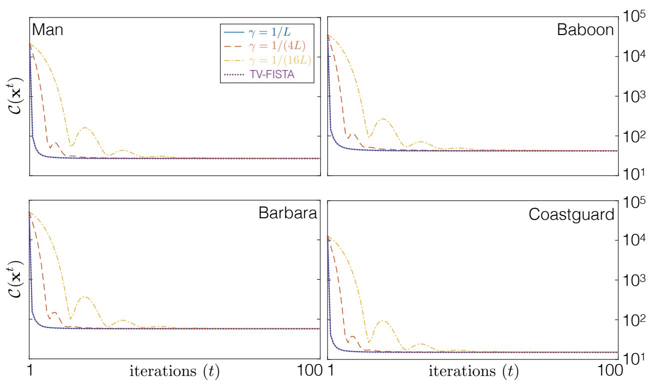

Fig. 4. Recovery of test images from blurry and noisy measurements. The TV cost functional $\mathcal{C}(\mathbf{x}^t)$ is plotted over 100 iterations for 3 distinct step-sizes γ. The dotted line plots the cost of the exact TV-FISTA. Note the nearly perfect match between FPPA at $\gamma = 1/L$ and TV-FISTA, which is consistent with their identical global convergence rates.

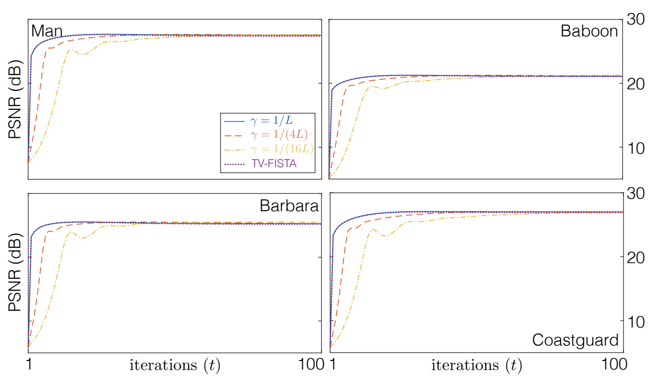

Fig. 5. Recovery of test images from blurry and noisy measurements. PSNR (dB) is plotted over 100 iterations for 3 distinct step-sizes γ. The dotted line plots the PSNR obtained by the exact TV-FISTA. The plot illustrates that the quality of the reconstructed images is approximately the same for both FPPA and TV-FISTA.

TV-FISTA was set to 100, also with a stopping criterion (24).

Figure 3 illustrates the evolution of the relative cost accuracy at every iteration of FPPA. It also reports the theoretical upper bound on the performance of FISTA in (11), as well as the actual evolution of the relative cost accuracy at every iteration of TV-FISTA. Figures 4 and 5 show the evolution of the cost $\mathcal{C}$ and PSNR, respectively, in the first 100 iterations of the algorithms. These figures highlight the convergence of FFPA and TV-FISTA within those 100 iterations, which indicates that the stoping criterion selected above was sufficiently strict. Finally, Figures 6 and 7 offer visual and quantitative evaluation of the final estimated images for *Man* and *Baboon*.

Proposition 1 suggests that the gap $(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*))$ is proportional to the step-size and shrinks to 0 as the step-size is reduced. Such behavior is clearly observed both in Table I and Figure 3. This suggests that our theoretical result is also valid in practice. On average, the relative cost accuracy for Set14 is about 1.33% at $\gamma = 1/L$, and decreases further for $\gamma = 1/(4L)$ and $\gamma = 1/(16L)$. Additionally, we note that the solution of our algorithm is very close to that of TV-FISTA visually and quantitatively. This implies that, while requiring no nested iterations, FPPA can potentially approximate the solution of TV with arbitrarily accurate precision at $\mathcal{O}(1/t^2)$ convergence rate of FISTA. Note also that FPPA is substantially faster that the standard approach that requires sub-iterations. For example in our simulations, FPPA achieved an average speed-up of 16 for $\gamma = 1/L$ on the Set14 images.

B. Discussion on TV-based Imaging

Minimization of TV regularized cost functionals is one of many approaches for reconstructing images from their linear measurements. A vast majority of these approaches rely on some form of prior information or constraints for regularizing the image formation process [39]. Depending on the type of prior information, algorithms can be loosely classified into several categories including traditional linear methods [40], Bayesian and statistical methods [35], [41]–[44], optimization based methods with pre-specified regularizers [45]–[48], patch based methods exploiting similarities in a given image [49]–[52], methods based on dictionary learning [53]–[55], supervised learning approaches based on deep convolutional networks (CNNs) [37], [38], [56]–[61].

Fig. 6. Recovery of *Man* from blurry and noisy measurements at 30 dB input SNR. The values in the top-right corner correspond to the PSNR in dB. (a) original; (b) measurements; (c) the TV-FISTA solution; (d) $\gamma = 1/L$; (e) $\gamma = 1/(4L)$; (f) $\gamma = 1/(16L)$. Even for $\gamma = 1/L$ the solution of the fast parallel proximal algorithm is visually and quantitatively close to the TV result. It takes about 37, 62, 96, and 639 seconds to obtain (c), (d), (e), and (f), respectively.

It has been widely reported that powerful patch-based methods based on the BM3D algorithm outperform TV on certain image restoration problems such as deblurring and denoising [47], [52]. Similar improvements were observed by another class of powerful methods based on deep convolutional networks [37], [57]. Nonetheless, each reconstruction approach has a distinct set of advantages and drawbacks that influences its applicability to various imaging problems. For example, patch-based methods rely on the block-matching procedure for grouping similar image patches. This implies that these methods require a suitable initial estimate of the image for a reliable block-matching, which makes them ideal for image denoising or deblurring [52], but makes their generalization to arbitrary imaging problems difficult. On the other hand, CNN based methods have a simple structure as a succession of convolutions. These methods, therefore, enjoy lower computational complexity for reconstruction compared to the patch-based methods. However, they typically require a separate training procedure over a sufficiently large image dataset. For example, Dong *et al.* [56] report that it took about three days to train their SR-CNN model on 24800 sub-images of size 32×32, extracted from 91 training images. Similarly, Chen and Pock [38] report 20.8 hours of training on 400 images of size 180×180. Additionally, model parameters in such imaging methods are highly optimized for a given problem, which implies that a slight modification in the acquisition system requires complete retraining of the network. Finally, while large training datasets are easy to generate for certain class of problems such as, for example, image super-resolution, they are harder to obtain for other applications such as bio-microscopy or medical imaging.

Compared to more advanced methods such as BM3D [52] or SR-CNN [37], TV based imaging does not rely on a suitable initialization for block-matching or require an additional training procedure. This makes it straightforward to apply to a larger set of imaging problems including image restoration [3], [15], depth imaging [62], [63], magnetic resonance imaging (MRI) [5], [64], computer tomography (CT) [33], phase-contrast tomography [65], optical microscopy [66]–[68], and inverse wave scattering [69]. In particular, TV imaging algorithms are particularly well-suited for very large-scale 3D imaging problems, where training and block-matching become prohibitively expensive. In such applications, it becomes crucial to have access to fast optimization algorithms for TV such as the one proposed here.

TV has been extensively compared in several prior works and a comprehensive comparison falls beyond the scope of this paper. Nonetheless, Table II summarizes its performance, in terms of PSNR (dB) and running time (sec), on image super-resolution over a dataset `Set14`. Specifically, we exactly reproduce the image upscaling problem that was also considered in previous works [37], [38], [54]. We report the results of both FPPA and TV-FISTA, as well as the results of a simple bicubic interpolation, Hessian Schatten-Norm Regularization (HS_2) [47], and SR-CNN [37]. FPPA, TV-FISTA, and HS_2

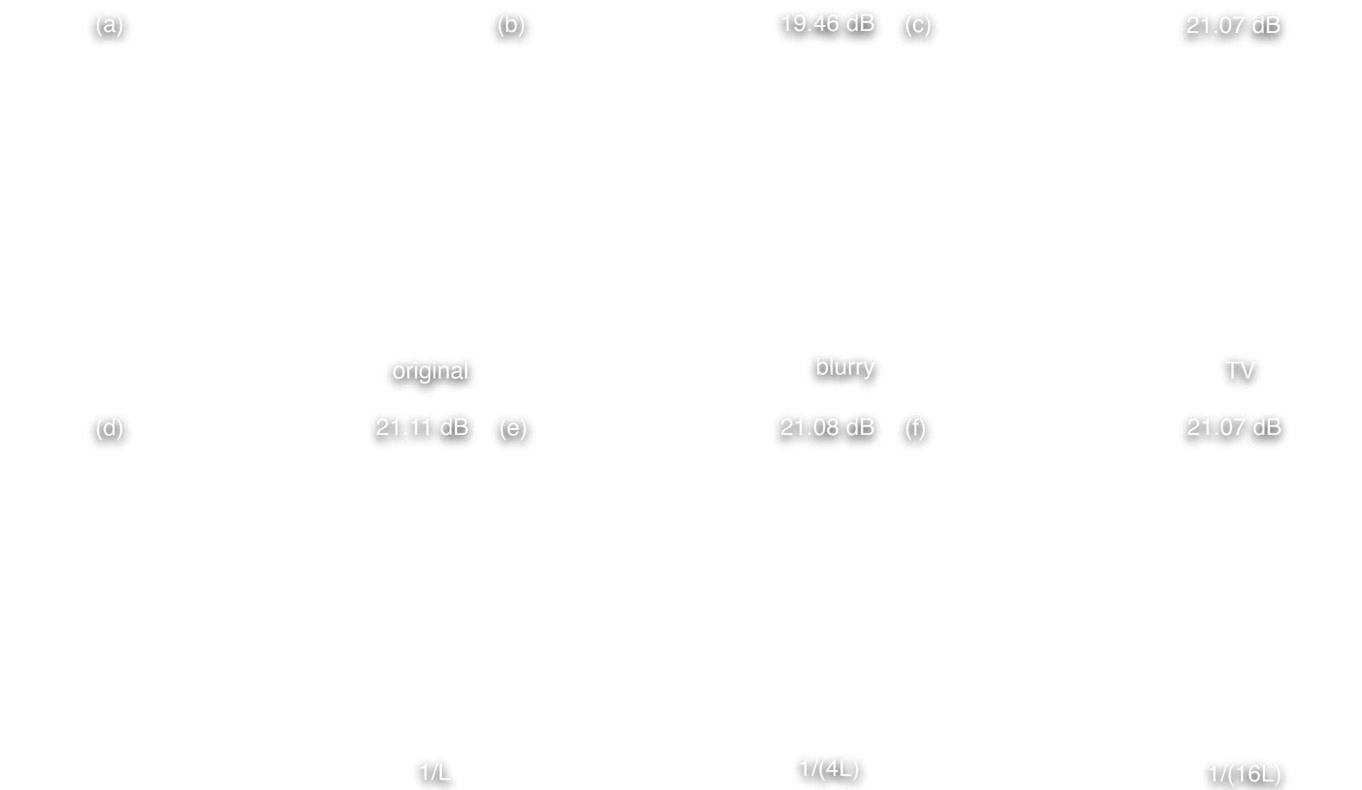

Fig. 7. Recovery of *Baboon* from blurry and noisy measurements at 30 dB input SNR. The values in the top-right corner correspond to the PSNR in dB. (a) original; (b) measurements; (c) the TV-FISTA solution; (d) $\gamma = 1/L$; (e) $\gamma = 1/(4L)$; (f) $\gamma = 1/(16L)$. Even for $\gamma = 1/L$ the solution of the fast parallel proximal algorithm is visually and quantitatively close to the TV result. It takes about 43, 72, 103, and 695 seconds to obtain (c), (d), (e), and (f), respectively.

were run for 20 iterations with PSNR optimal regularization parameters. The computation of the proximals of TV-FISTA and HS_2 was limited to 5 sub-iterations. We relied on the MATLAB implementations of HS_2 and SR-CNN that was provided by the authors. Since none of the methods were optimized for speed, the running times are expected to further improve after a careful code optimization.

The very first observation is that FPPA closely approximates the TV-FISTA solution at the fraction of the running time (0.04 dB difference for about $\times 3$ reduction in time). Additionally, both TV methods yield images that are within 0.4 dB compared to the powerful SR-CNN. As TV does not require an extensive model training procedure, this indicates that it can be a simpler, but an effective, alternative to SR-CNN when training is not practical or possible.

V. Conclusion

The fast parallel proximal method, which was presented in this paper, is beneficial in the context of anisotropic TV regularized image reconstruction, especially when the computation of the TV proximal is costly. We presented a mixture of theoretical and empirical evidence demonstrating that the method can accurately approximate 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 to isotropic variant of TV and by applying the methods to practical large scale imaging problems. Additionally, it would be beneficial to study the method when γ_t is decreased progressively, which could mitigate the stalling effect when γ is fixed to a small value.

VI. Appendix

A. Review of Convex Analysis

Before embarking on the actual proof of Proposition 1, it is convenient to summarize a few facts that will be used next.

A subgradient of a convex function $\mathcal{C}$ at $\mathbf{x}$ is any vector $\tilde{\nabla}\mathcal{C}(\mathbf{x})$ that satisfies the inequality

$$\mathcal{C}(\mathbf{y}) \geq \mathcal{C}(\mathbf{x}) + \langle \tilde{\nabla}\mathcal{C}(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \tag{25}$$

for all $\mathbf{y}$. When $\mathcal{C}$ is differentiable, the only possible choice for $\tilde{\nabla}\mathcal{C}$ is the gradient $\nabla\mathcal{C}$. The set of subgradients of $\mathcal{C}$ at $\mathbf{x}$ is the subdifferential of $\mathcal{C}$ at $\mathbf{x}$, denoted $\partial\mathcal{C}(\mathbf{x})$. The condition that $\tilde{\nabla}\mathcal{C}$ be a subgradient of $\mathcal{C}$ at $\mathbf{x}$ can then be written $\tilde{\nabla}\mathcal{C}(\mathbf{x}) \in \partial\mathcal{C}(\mathbf{x})$.

We also remind another fundamental property of a smooth and continuously differentiable function $\mathcal{D}$ with a Lipschitz continuous gradient and Lipschitz constant L. For any $\gamma \in (0, 1/L]$, such functions satisfy

$$\mathcal{D}(\mathbf{x}) \leq \mathcal{D}(\mathbf{y}) + \langle \nabla D(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{y}\|_{\ell_2}^2, \tag{26}$$

and all $\mathbf{x}, \mathbf{y}$

TABLE II
UPSCALING BY FACTOR $\times 3$ PERFORMANCE IN TERMS OF PSNR AND RUNTIME FOR THE IMAGES FROM SET14.

Set14	Bicubic		FPPA		TV		HS_2		SR-CNN	
Images	PSNR (dB)	Time (s)	PSNR (dB)	Time (s)	PSNR (dB)	Time (s)	PSNR (dB)	Time (s)	PSNR (dB)	Time (s)
Baboon	23.21	0.004	23.53	0.533	23.53	1.760	23.54	4.079	23.60	3.874
Barbara	26.25	0.004	26.77	0.918	26.77	3.606	26.75	8.098	26.66	6.744
Bridge	24.40	0.003	24.94	0.556	24.94	1.946	24.96	4.340	25.07	4.219
Coastguard	26.55	0.003	27.03	0.233	27.03	0.659	27.04	1.623	27.20	1.757
Comic	23.12	0.003	23.92	0.233	23.91	0.604	23.90	1.426	24.39	1.694
Face	32.82	0.003	33.35	0.201	33.38	0.501	33.59	1.202	33.58	1.505
Flowers	27.23	0.004	28.53	0.370	28.37	1.800	28.36	2.842	28.97	2.763
Foreman	31.18	0.003	33.17	0.222	33.05	0.679	32.84	1.642	33.35	1.742
Lenna	31.68	0.004	32.73	0.564	32.69	2.032	32.94	4.335	33.39	4.057
Man	27.01	0.003	27.84	0.714	27.76	1.977	27.78	4.469	28.18	4.223
Monarch	29.43	0.005	31.71	0.881	31.54	3.232	30.75	7.729	32.39	6.529
Pepper	32.39	0.005	33.91	0.532	33.69	1.962	33.55	4.465	34.35	4.353
Ppt3	23.71	0.004	25.21	0.703	25.57	2.758	24.86	6.226	26.02	5.531
Zebra	26.63	0.003	28.36	0.485	28.22	1.765	27.95	3.911	28.87	3.848
Average	**27.54**	**0.004**	**28.64**	**0.510**	**28.60**	**1.806**	**28.49**	**4.028**	**29.00**	**3.774**

The proximal operator is defined as

$$\mathbf{x} = \mathrm{prox}_{\gamma\mathcal{R}}(\mathbf{z}) \tag{27a}$$

$$\arg\min_{\mathbf{x}\in\mathbb{R}^N} \left\{\frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_{\ell_2}^2 + \gamma\mathcal{R}(\mathbf{x})\right\} \tag{27b}$$

where $\gamma > 0$ and $\mathcal{R}$ is a convex continuous function. The proximal operator is characterized by the following inclusion, for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$

$$\mathbf{x} = \mathrm{prox}_{\gamma\mathcal{R}}(\mathbf{z}) \Leftrightarrow \frac{\mathbf{z}-\mathbf{x}}{\gamma} \in \partial\mathcal{R}(\mathbf{x}). \tag{28}$$

B. Proof of Proposition 1

We consider the following algorithm, which is perfectly equivalent to the fast parallel proximal algorithm (14)

$$\mathbf{u}^t \leftarrow (1-1/q_t)\mathbf{x}^{t-1} + (1/q_t)\mathbf{v}^{t-1} \tag{29a}$$

$$\mathbf{x}^t \leftarrow \frac{1}{K}\sum_{k=1}^{K} \mathrm{prox}_{\gamma\mathcal{R}_k}(\mathbf{u}^t - \gamma\nabla\mathcal{D}(\mathbf{u}^t)) \tag{29b}$$

$$\mathbf{v}^t \leftarrow \mathbf{x}^{t-1} + q_t(\mathbf{x}^t - \mathbf{x}^{t-1}), \tag{29c}$$

where $\mathbf{x}^0 = \mathbf{v}^0$, $q_0 = 1$, and q_t satisfies

$$q_t^2 - q_t \leq q_{t-1}^2, \tag{30}$$

for all $t = 1, 2, \dots$. To see the equivalence of (29) to the fast parallel proximal algorithm (14), first set q_t as in (14b) and then eliminate the auxiliary variables $\{\mathbf{v}^t\}$ by plugging (29c) into (29a).

We start by using (26) to find an upper bound for $\mathcal{D}$ at $\mathbf{x}^t$

$$\mathcal{D}(\mathbf{x}^t) \leq \mathcal{D}(\mathbf{u}^t) + \langle\nabla D(\mathbf{u}^t), \mathbf{x}^t - \mathbf{u}^t\rangle + \frac{1}{2\gamma}\|\mathbf{x}^t - \mathbf{u}^t\|_{\ell_2}^2. \tag{31}$$

We then define an intermediate quantity $\mathbf{x}_k^t \triangleq \mathrm{prox}_{\gamma\mathcal{R}_k}(\mathbf{u}^{t-1} - \gamma\nabla\mathcal{D}(\mathbf{u}^{t-1}))$. The optimality conditions for (29b) imply that there must equst K subgradient vectors $\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t) \in \partial\mathcal{R}_k(\mathbf{x}_k^t)$ such that

$$\mathbf{x}_k^t = \mathbf{u}^t - \gamma(\nabla\mathcal{D}(\mathbf{u}^t) + \tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t)). \tag{32}$$

This implies that

$$\mathbf{x}^t = \mathbf{u}^t - \gamma(\nabla\mathcal{D}(\mathbf{u}^t) + \mathbf{g}^t), \tag{33}$$

where

$$\mathbf{g}^t \triangleq \frac{1}{K}\sum_{k=1}^{K}\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t). \tag{34}$$

The relationships (32) and (33) together with bounds on the subgradients implies that

$$\|\mathbf{x}^t - \mathbf{x}_k^t\|_{\ell_2} = \|\gamma(\mathbf{g}^t - \tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t))\|_{\ell_2} \leq 2\gamma G. \tag{35}$$

We then bound $\mathcal{R}_k$ at any $\mathbf{z} \in \mathbb{R}^N$ as follows

$$\begin{aligned}
\mathcal{R}_k(\mathbf{z}) &\overset{\text{(a)}}{\geq} \mathcal{R}_k(\mathbf{x}_k^t) + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{z} - \mathbf{x}_k^t\rangle && (36a)\\
&= \mathcal{R}_k(\mathbf{x}_k^t) + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{z} - \mathbf{x}^t\rangle && (36b)\\
&\quad + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{x}^t - \mathbf{x}_k^t\rangle\\
&\overset{\text{(b)}}{\geq} \mathcal{R}_k(\mathbf{x}^t) + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}^t), \mathbf{x}_k^t - \mathbf{x}^t\rangle && (36c)\\
&\quad + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{z} - \mathbf{x}^t\rangle + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{x}^t - \mathbf{x}_k^t\rangle\\
&= \mathcal{R}_k(\mathbf{x}^t) + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{z} - \mathbf{x}^t\rangle && (36d)\\
&\quad + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t) - \tilde{\nabla}\mathcal{R}_k(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}_k^t\rangle,
\end{aligned}$$

where in (a) and (b) we used the convexity of $\mathcal{R}_k$. By rearranging (36), we obtain for any $\mathbf{z} \in \mathbb{R}^N$

$$\begin{aligned}
\mathcal{R}_k(\mathbf{x}^t) &\leq \mathcal{R}_k(\mathbf{z}) - \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{z} - \mathbf{x}^t\rangle && (37a)\\
&\quad + \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t) - \tilde{\nabla}\mathcal{R}_k(\mathbf{x}^t), \mathbf{x}_k^t - \mathbf{x}^t\rangle\\
&\overset{\text{(a)}}{\leq} \mathcal{R}_k(\mathbf{z}) - \langle\tilde{\nabla}\mathcal{R}_k(\mathbf{x}_k^t), \mathbf{z} - \mathbf{x}^t\rangle + 4\gamma G^2, && (37b)
\end{aligned}$$

where in (a) we used Cauchy-Schwarz inequality followed by (35). By averaging (37) over k and using (33), we obtain

$$\begin{aligned}
&\mathcal{R}(\mathbf{x}^t) && (38)\\
&\quad\leq \mathcal{R}(\mathbf{z}) + \frac{1}{\gamma}\langle\mathbf{u}^t - \gamma\nabla\mathcal{D}(\mathbf{u}^t) - \mathbf{x}^t, \mathbf{x}^t - \mathbf{z}\rangle + 4\gamma G^2,
\end{aligned}$$

for any $\mathbf{z} \in \mathbb{R}^N$. We next add bounds (31) and (38) and use the convexity of $\mathcal{D}$ to obtain

$$\mathcal{C}(\mathbf{x}^t) \leq \mathcal{C}(\mathbf{z}) + \frac{1}{\gamma}\langle \mathbf{x}^t - \mathbf{u}^t, \mathbf{z} - \mathbf{x}^t \rangle + \frac{1}{2\gamma}\|\mathbf{x}^t - \mathbf{u}^t\|_{\ell_2}^2 + 4\gamma G^2, \tag{39}$$

for all $\mathbf{z} \in \mathbb{R}^N$. By evaluating (39) at $\mathbf{z} = \mathbf{x}^{t-1}$ and $\mathbf{z} = \mathbf{x}^*$ and taking the convex combination of the bounds, we obtain

$$\begin{aligned} &\mathcal{C}(\mathbf{x}^t) - (1 - 1/q_t)\mathcal{C}(\mathbf{x}^{t-1}) - (1/q_t)\mathcal{C}(\mathbf{x}^*) && (40) \\ &= [\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*)] - (1 - 1/q_t)[\mathcal{C}(\mathbf{x}^{t-1} - \mathcal{C}(\mathbf{x}^*)) \\ &\leq \frac{1}{\gamma}\langle \mathbf{x}^t - \mathbf{u}^t, \frac{1}{q_t}\mathbf{x}^* + \left(1 - \frac{1}{q_t}\right)\mathbf{x}^{t-1} - \mathbf{x}^t \rangle \\ &\quad + \frac{1}{\gamma}\|\mathbf{x}^t - \mathbf{u}^t\|_{\ell_2}^2 + 4\gamma G^2 \\ &= \frac{1}{2\gamma q_t^2}\left(\|\mathbf{v}^{t-1} - \mathbf{x}^*\|_{\ell_2}^2 - \|\mathbf{v}^t - \mathbf{x}^*\|_{\ell_2}^2\right) + 4\gamma G^2, && (41) \end{aligned}$$

where in the last step we completed the squares and used the definition of the auxiliary variables $\{\mathbf{v}^t\}$ in (29a) and (29c). We thus get the following recursive relationship

$$\begin{aligned} &\gamma q_t^2(\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*)) + \frac{1}{2}\|\mathbf{v}^t - \mathbf{x}^*\|_{\ell_2}^2 && (42) \\ &\quad \leq \gamma(q_t^2 - q_t)(\mathcal{C}(\mathbf{x}^{t-1}) - \mathcal{C}(\mathbf{x}^*)) + \frac{1}{2}\|\mathbf{v}^{t-1} - \mathbf{x}^*\|_{\ell_2}^2 \\ &\quad\quad + 4q_t^2\gamma^2 G^2. && (43) \end{aligned}$$

By using the bound (30), using a particular $q_t = (t+1)/2$, and iterating over t, we get

$$\mathcal{C}(\mathbf{x}^t) - \mathcal{C}(\mathbf{x}^*) \leq \frac{2}{\gamma(t+1)^2}\|\mathbf{x}^0 - \mathbf{x}^*\|_{\ell_2}^2 + 4\gamma G^2. \tag{44}$$

This completes the proof.

ACKNOWLEDGEMENTS

The author would like to thank H. Mansour and A. Vetro for their help in the preparation of this manuscript.

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., vol. 3024, New York, 2004, pp. 1–13.

[12] 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.

[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] ——, "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.

[16] 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.

[17] A. d'Aspremont, "Smooth optimization with approximate gradient," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1171–1183, 2008.

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

[19] M. Schmidt, N. Le Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Advances in Neural Information Processing Systems 24*, Granada, Spain, December 12-15, 2011.

[20] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Math. Program. Ser. A*, vol. 146, no. 1-2, pp. 37–75, 2013.

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

[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] ——, "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.

[24] L. Condat, "A direct algorithm for 1-D total variation denoising," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1054–1057, November 2013.

[25] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. 20th Int. Conf. Machine Learning (ICML)*, Washington DC, USA, August 21-24, 2003.

[26] J. Duchi and Y. Singer, "Efficient online and batch learning using forward backward splitting," *J. Mach. Learn. Res*, vol. 10, pp. 2899–2934, 2009.

[27] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, pp. 543–547, 1983, (in Russian).

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

[29] A. K. Fletcher, K. Ramchandran, and V. K. Goyal, "Wavelet denoising by recursive cycle spinning," in *Proc. IEEE Int. Conf. Image Proc. (ICIP'02)*, Rochester, NY, USA, September 22–25, 2002, pp. II.873–II.876.

[30] C. Vonesch and M. Unser, "A fast thresholded Landweber algorithm for wavelet-regularized multidimensional deconvolution," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 539–549, April 2008.

[31] ——, "A fast multilevel algorithm for wavelet-regularized image restoration," *IEEE Trans. Image Process.*, vol. 18, no. 3, pp. 509–523, March 2009.

[32] M. Guerquin-Kern, M. Häberlin, K. P. Prüssmann, and M. Unser, "A fast wavelet-based reconstruction method for magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 30, no. 9, pp. 1649–1660, September 2011.

[33] S. Ramani and J. A. Fessler, "A hybrid regularizer combining orthonormal wavelets and finite differences for statistical reconstruction in 3-D CT," in *Proc. 2nd Intl. Mtg. on image formation in X-ray CT*, Salt Lake City, UT, USA, 2012, pp. 348–351.

[34] 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, 2016, pp. 4697–4701.

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

[36] R. Timofte, V. De Smet, and L. Van Gool, “Anchored neighborhood regression for fast example-based super-resolution,” in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, Sydney, Australia, Dec 1-8, 2013, pp. 1920–1927.

[37] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Trans. Patt. Anal. and Machine Intell.*, vol. 38, no. 2, pp. 295–307, February 2016.

[38] Y. Chen and T. Pock, “Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration,” 2016, arXiv:1508.02848 [cs.CV].

[39] A. Ribés and F. Schmitt, “Linear inverse problems in imaging,” *IEEE Signal Process. Mag.*, vol. 25, no. 4, pp. 84–99, July 2008.

[40] A. N. Tikhonov and V. Y. Arsening, *Solution of Ill-Posed Problems*. Winston-Wiley, 1977.

[41] E. P. Simoncelli, *Bayesian denoising of visual images in the wavelet domain*. Springer Verlag, 1999, vol. 141.

[42] M. Nikolova, “Model distortions in Bayesian MAP reconstruction,” *Inverse Probl. Imag.*, vol. 1, no. 2, pp. 399–422, May 2007.

[43] M. Seeger, “Bayesian inference and optimal design for the sparse linear model,” *J. Machine Learning Research*, vol. 9, pp. 759–813, September 2008.

[44] U. S. Kamilov, “Sparsity-driven statistical inference for inverse problems,” EPFL Thesis no. 6545 (2015), 198 p., Swiss Federal Institute of Technology Lausanne (EPFL), March 27, 2015.

[45] M. Belge, M. E. Kilmer, and E. L. Miller, “Wavelet domain image restoration with adaptive edge-preserving regularization,” *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 597–608, April 2000.

[46] L. Sendur and I. W. Selesnick, “Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency,” *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2744–2756, 2002.

[47] S. Lefkimmiatis, A. Bourquard, and M. Unser, “Hessian-based norm regularization for image restoration with biomedical applications,” *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 983–995, March 2012.

[48] E. Bostan, U. S. Kamilov, M. Nilchian, and M. Unser, “Sparse stochastic processes and discretization of linear inverse problems,” *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2699–2710, July 2013.

[49] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Trans. Image Process.*, vol. 16, no. 16, pp. 2080–2095, August 2007.

[50] ——, “Image restoration by sparse 3D transform-domain collaborative filtering,” in *Proc. SPIE Elctron. Imag.*, vol. 6812, San Jose, CA, January 2008, p. 681207.

[51] A. Buades, B. Coll, and J. M. Morel, “Image denoising methods. A new nonlocal principle.” *SIAM Rev*, vol. 52, no. 1, pp. 113–147, 2010.

[52] A. Danielyan, V. Katkovnik, and K. Egiazarian, “BM3D frames and variational image deblurring,” *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1715–1728, April 2012.

[53] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Novemberg 2006.

[54] R. Zeyde, M. Elad, and M. Protter, *Curves and Surfaces*. Springer, 2012, ch. On single image scale-up using sparse-representations, pp. 711–730.

[55] J. Mairal, F. Bach, and J. Ponce, “Sparse modeling for image and vision processing,” *Foundations and Trends in Machine Learning*, vol. 8, no. 2-3, pp. 1–199, 2014.

[56] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in *Proc. ECCV*, Zurich, Switzerland, September 6-12, 2014, pp. 184–199.

[57] U. Schmidt and S. Roth, “Shrinkage fields for effective image restoration,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, June 23-28, 2014, pp. 2774–2781.

[58] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, “Deep networks for image super-resolution with sparse prior,” in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, Santiago, Chile, December 13-16, 2015, pp. 370–378.

[59] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, “Convolutional sparse coding for image super-resolution,” in *Proc. IEEE Int. Conf. Comp. Vis. (ICCV)*, Santiago, Chile, December 13-16, 2015, pp. 1823–1831.

[60] Y. Chen, W. Yu, and T. Pock, “On learning optimized reaction diffuction processes for effective image restoration,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 8-10, 2015, pp. 5261–5269.

[61] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 26-July 1, 2016.

[62] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Sydney, NSW, Australia, December 1-8, 2013, pp. 993–1000.

[63] J. Castorena, U. S. Kamilov, and P. T. Boufounos, “Autocalibration of LIDAR and optical cameras via edge alignment,” in *IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP 2016)*, Shanghai, China, March 20–25 2016, pp. 2862–2866.

[64] E. J. Candès, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *J. of Fourier Anal. Appl.*, vol. 14, no. 5–6, pp. 877–905, October 2008.

[65] M. Nilchian and M. Unser, “Differential phase-contrast X-ray computed tomography: From model discretization to image reconstruction,” in *Proc. Int. Symp. Biomedical Imaging*, Barcelona, Spain, May 2012.

[66] Y. Sung and R. R. Dasari, “Deterministic regularization of three-dimensional optical diffraction tomography,” *J. Opt. Soc. Am. A*, vol. 28, no. 8, pp. 1554–1561, August 2011.

[67] J. W. Lim, K. R. Lee, K. H. Jin, S. Shin, S. E. Lee, Y. K. Park, and J. C. Ye, “Comparative study of iterative reconstruction algorithms for missing cone problems in optical diffraction tomography,” *Opt. Express*, vol. 23, no. 13, pp. 16 933–16 948, June 2015.

[68] 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.

[69] U. S. Kamilov, D. Liu, H. Mansour, and P. T. Boufounos, “A recursive Born approach to nonlinear inverse scattering,” *IEEE Signal Process. Lett.*, vol. 23, no. 8, pp. 1052–1056, August 2016.

Ulugbek S. Kamilov (S’11–M’15) is a Research Scientist in the Computational Sensing team at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. Dr. Kamilov obtained his B.Sc. and M.Sc. in Communication Systems, and Ph.D. in Electrical Engineering from the École polytechnique fédérale de Lausanne (EPFL), Switzerland, in 2008, 2011, and 2015, respectively. In 2007, he was an Exchange Student at Carnegie Mellon University (CMU), Pittsburgh, PA, USA, in 2010, a Visiting Student at Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, and in 2013, a Visiting Student Researcher at Stanford University, Stanford, CA, USA.

Dr. Kamilov’s research focus is computational imaging with an emphasis on the development and analysis of large-scale computational techniques for biomedical and industrial applications. His research interests cover multimodal sensor fusion, tomographic imaging, machine learning, through-the-wall imaging, and distributed radar sensing. He has co-authored 16 journal and 28 conference publications in these areas. His Ph.D. thesis work on Learning Tomography (LT) was selected as a finalist for EPFL Doctorate Awards 2016 and was featured in the “News and Views” of Nature. Since 2016, Dr. Kamilov is a member IEEE Special Interest Group on Computational Imaging.