Projected Nesterov's Proximal-Gradient Algorithm for Sparse Signal Recovery

Renliang Gu and Aleksandar Dogandžić, *Senior Member, IEEE*

***Abstract*—We develop a projected Nesterov's proximal-gradient (PNPG) approach for sparse signal reconstruction that combines adaptive step size with Nesterov's momentum acceleration. The objective function that we wish to minimize is the sum of a convex differentiable data-fidelity (negative log-likelihood (NLL)) term and a convex regularization term. We apply sparse signal regularization where the signal belongs to a closed convex set within the closure of the domain of the NLL; the convex-set constraint facilitates flexible NLL domains and accurate signal recovery. Signal sparsity is imposed using the ℓ_1-norm penalty on the signal's linear transform coefficients. The PNPG approach employs a projected Nesterov's acceleration step with restart and a duality-based inner iteration to compute the proximal mapping. We propose an *adaptive* step-size selection scheme to obtain a good local majorizing function of the NLL and reduce the time spent backtracking. Thanks to step-size adaptation, PNPG converges faster than the methods that do not adjust to the local curvature of the NLL. We present an integrated derivation of the momentum acceleration and proofs of $\mathcal{O}(k^{-2})$ objective function convergence rate and convergence of the iterates, which account for adaptive step size, inexactness of the iterative proximal mapping, and the convex-set constraint. The tuning of PNPG is largely application independent. Tomographic and compressed-sensing reconstruction experiments with Poisson generalized linear and Gaussian linear measurement models demonstrate the performance of the proposed approach.**

***Index Terms*—Convex optimization, Nesterov's momentum acceleration, sparse signal reconstruction, Poisson compressed sensing, proximal-gradient methods.**

I. Introduction

MOST natural signals are well described by only a few significant coefficients in an appropriate transform domain, with the number of significant coefficients much smaller than the signal size. Therefore, for a vector $\boldsymbol{x} \in \mathbb{R}^p$ that represents the signal and an appropriate *sparsifying dictionary* matrix Ψ, $\Psi^H \boldsymbol{x}$ is a signal transform-coefficient vector with most elements having negligible magnitudes. Real-valued $\Psi \in \mathbb{R}^{p\times p'}$ can accommodate discrete wavelet transform (DWT) or gradient-map sparsity with anisotropic total-variation (TV) sparsifying transform (with $\Psi = [\Psi_{\rm v}\ \Psi_{\rm h}]$); a complex-valued $\Psi = \Psi_{\rm v} + {\rm j}\Psi_{\rm h} \in \mathbb{C}^{p\times p'}$ can accommodate gradient-map sparsity and the 2D isotropic TV sparsifying transform; here $\Psi_{\rm v}, \Psi_{\rm h} \in \mathbb{R}^{p\times p'}$ are the vertical and horizontal difference matrices similar to those in [1, Sec. 15.3.3]. The idea behind compressed sensing [2] is to *sense* the significant components of $\Psi^H \boldsymbol{x}$ using a small number of measurements; here, "^{H}" denotes the conjugate transpose.

We use the negative log-likelihood (NLL) (data-fidelity) function $\mathcal{L}(\boldsymbol{x})$ to describe the noisy measurement process. Consider signals $\boldsymbol{x}$ that belong to a closed convex set C and assume

$$C \subseteq {\rm cl}({\rm dom}\ \mathcal{L}) \tag{1}$$

which ensures that $\mathcal{L}(\boldsymbol{x})$ is computable for all $\boldsymbol{x} \in {\rm int}\, C$. If $C \setminus {\rm dom}\ \mathcal{L}$ is not empty, then $\mathcal{L}(\boldsymbol{x})$ is not computable in it, which needs special attention; see Section III. The nonnegative signal scenario with

$$C = \mathbb{R}^p_+ \tag{2}$$

is of significant practical interest and applicable to X-ray computed tomography (CT), single photon emission computed tomography (SPECT), positron emission tomography (PET), and magnetic resonance imaging (MRI), where the pixel values correspond to inherently nonnegative density or concentration maps [3]. Harmany *et al.* consider such a nonnegative sparse signal model and develop in [4] and [5] a convex-relaxation sparse Poisson-intensity reconstruction algorithm (SPIRAL) and a linearly constrained gradient projection method for Poisson and Gaussian linear measurements, respectively. In addition to signal nonnegativity, other convex-set constraints have been considered in the literature: prescribed value in the Fourier domain; box, geometric, and total-energy constraints; intersections of these sets [6]; and unit simplex [7].

We adopt the analysis regularization framework and minimize

$$f(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) + ur(\boldsymbol{x}) \tag{3a}$$

with respect to the signal $\boldsymbol{x}$, where $\mathcal{L}(\boldsymbol{x})$ is a differentiable convex NLL and

$$r(\boldsymbol{x}) = I_C(\boldsymbol{x}) + \rho(\boldsymbol{x}) \tag{3b}$$

is a convex regularization term that imposes convex-set constraint on $\boldsymbol{x}$, $\boldsymbol{x} \in C$, and sparsity of an appropriate transformed $\boldsymbol{x}$ through the convex penalty $\rho(\boldsymbol{x})$ [4], [8]–[11]. Here, $u > 0$

Manuscript received July 1, 2016; revised January 18, 2017 and March 22, 2017; accepted March 24, 2017. Date of publication April 6, 2017; date of current version May 5, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Cédric Févotte. This work was supported by the U.S. National Science Foundation under Grant CCF-1421480. This paper was presented in part at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, November 2014, and in part at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, November 2015. *(Corresponding author: Renliang Gu.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: renliang@iastate.edu; ald@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2017.2691661

is a scalar tuning constant that quantifies the weight of the regularization term, and $I_C(\boldsymbol{x}) \triangleq \begin{cases} 0, & \boldsymbol{x} \in C \\ +\infty, & \text{otherwise} \end{cases}$ is the indicator function. The penalty $\rho(\boldsymbol{x})$ is often selected as the ℓ_1-norm of the signal transform-coefficient vector [11]:

$$\rho(\boldsymbol{x}) = \|\Psi^H \boldsymbol{x}\|_1. \tag{4}$$

Define the proximal operator for a function $r(\boldsymbol{x})$ scaled by $\lambda > 0$ at argument $\boldsymbol{a} \in \mathbb{R}^p$:

$$\text{prox}_{\lambda r}\boldsymbol{a} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^p} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{a}\|_2^2 + \lambda r(\boldsymbol{x}). \tag{5}$$

In this paper (see also [8], [9]), we develop a projected Nesterov's proximal-gradient (PNPG) method whose momentum acceleration accommodates adaptive step-size selection and convex-set constraint on the signal $\boldsymbol{x}$. Computing the proximal operator with respect to $r(\boldsymbol{x})$ in (3b) needs iteration and is therefore inexact [12]–[14]. We establish conditions for the $\mathcal{O}(k^{-2})$ convergence rate of the objective function as well as the convergence of PNPG iterates. These results are the first for an accelerated proximal-gradient (PG) method *with step-size adaptation* (and, therefore, adjustment to the local curvature of the NLL) that

- establish convergence of the iterates (Theorem 2) and
- incorporate inexact proximal operators into objective function convergence rate and convergence of the iterates analyses (Theorems 1 and 2).

We modify the original Nesterov's acceleration [15], [16] so that we can establish these results when the step size is adaptive and adjusts to the local curvature of the NLL. (Local-curvature adjustments of the NLL by step-size adaptation have also been used in other algorithms under different contexts in [17]–[19]; see also the following and discussion in Section IV-A1.) Our integration of the adaptive step size and convex-set constraint extends the application of the Nesterov-type acceleration to more general measurement models than those used previously, such as the Poisson compressed-sensing scenario described in Section II-A. Furthermore, a convex-set constraint can bring significant improvement to signal reconstructions compared with imposing signal sparsity only, as illustrated in Section V-B. See Section IV-A for further discussion of $\mathcal{O}(k^{-2})$ acceleration approaches [10], [16], [18], [20], [21].

Optimization problems (3a) with composite penalty-term structure in (3b) have been considered in [4], [12], [22], [23], which use PG (forward-backward)–type methods with nested inner iterations. The general optimization approach in these references is close to ours. Unlike PNPG, these methods approximate the NLLs whose gradients are not Lipschitz continuous and [4], [12], [22] do not have fast $\mathcal{O}(k^{-2})$ convergence-rate guarantees; [22] observes the benefits of larger step sizes and step-size adaptation. The nested forward-backward splitting iteration in [23] applies fast iterative shrinkage-thresholding algorithm (FISTA) [24] in both the outer and inner loops using duality to formulate the inner iteration; however, it does not employ step-size adaptation or analyze effects of inexact proximal-mapping computations. References [11], [23], [25]–[29] describe splitting schemes to minimize (3a), where [11], [26] are inspired by the parallel proximal algorithm (PPXA) [30]. Some splitting schemes, e.g., [11], [26], apply proximal operations on individual summands $\mathcal{L}(\boldsymbol{x})$, $u\rho(\boldsymbol{x})$, and $I_C(\boldsymbol{x})$, which is useful if all individual proximal operators are easy to compute. Both [11] and generalized forward-backward (GFB) splitting [25] require inner iterations to solve $\text{prox}_{\lambda\rho}\boldsymbol{a}$ for $\rho(\boldsymbol{x})$ in (4) in the general case where the sparsifying matrix Ψ is not orthogonal. Reference [23] applies the primal-dual approach by Chambolle and Pock [27], which allows solving its Poisson reconstruction problems without approximating the NLL: (3a) is split into $\mathcal{L}(\boldsymbol{x})$ and $r(\boldsymbol{x})$ and also into $\mathcal{L}(\boldsymbol{x}) + \rho(\boldsymbol{x})$ and $I_C(\boldsymbol{x})$, where the second approach (termed CP) does not require nested iterations. The primal-dual splitting (PDS) method in [28], [29] does not require inner iterations for general $\mathcal{L}(\boldsymbol{x})$ and sparsifying matrix. GFB and PDS need Lipschitz-continuous gradient of $\mathcal{L}$ and the value of the Lipschitz constant is important for tuning their parameters. The convergence rate of both GFB and PDS methods can be upper-bounded by C/k where k is the number of iterations and the constant C is determined by values of the tuning proximal and relaxation constants [31], [32]. In Section V, we show the performances of CP, GFB, and PDS.

Variable-metric methods with *problem-specific* diagonal scaling matrices have been considered in [10], [19], [33]; [19] applies Barzilai-Borwein (BB) step size and an Armijo line search for the overrelaxation parameter. It accounts for inexact proximal operator and establishes convergence of iterates but *does not* employ acceleration or provide fast convergence-rate guarantees. [19] does not require the Lipschitz continuity of the gradient of the NLL in general, except for proving the convergence rate of the objective function. Salzo [33] analyzes variable-metric algorithms without acceleration (of the type [19]) and relies on the uniform continuity of $\nabla\mathcal{L}$ for the convergence analysis of both objective function and iterates; however, [33] does not account for inexact proximal operators. In practice, special care is needed in selecting a good scaling matrix, and no clear guidelines are given in [10], [19], [33] for this selection. Setting the overrelaxation parameter in [33] to unity leads to a variable-metric/scaling scheme with an adaptive step size; further, setting the scaling matrix to identity leads to a PG iteration with adaptive step size.

Similar to templates for first-order conic solvers (TFOCS) [18], PNPG code is easy to maintain: for example, the proximal-mapping computation can be easily replaced as a module by the latest state-of-the-art solver. Furthermore, PNPG requires minimal *application-independent tuning*; indeed, we use the same set of tuning parameters in two different application examples. This is in contrast with the existing splitting methods, which require problem-dependent (NLL- and u-dependent) tuning, with convergence speed sensitive to the choice of tuning constants.

We review the notation: $\mathbf{0}$, $\mathbf{1}$, I, denoting the vectors of zeros and ones and identity matrix, respectively; "$\succeq$" is the elementwise version of "$\geq$"; "^{T}" and "^{H}" are transpose and conjugate transpose, respectively. For a vector $\boldsymbol{a} = (a_i)_{i=1}^N \in \mathbb{R}^N$, define the projection and soft-thresholding operators:

$$P_C(\boldsymbol{a}) = \arg\min_{\boldsymbol{x}\in C} \|\boldsymbol{x} - \boldsymbol{a}\|_2^2 \tag{6a}$$

$$[\mathcal{T}_\lambda(\boldsymbol{a})]_i = \text{sgn}(a_i)\max(|a_i| - \lambda, 0) \tag{6b}$$

and the elementwise exponential function $[\exp_\circ \boldsymbol{a}]_i = \exp a_i$. The projection onto $\mathbb{R}_+^N$ and the proximal operator (5) for the ℓ_1-norm $\|\boldsymbol{x}\|_1$ can be computed in closed form:

$$\left[P_{\mathbb{R}_+^N}(\boldsymbol{a})\right]_i = \max(a_i, 0), \qquad \operatorname{prox}_{\lambda\|\cdot\|_1}\boldsymbol{a} = \mathcal{T}_\lambda(\boldsymbol{a}). \tag{6c}$$

A. Preliminary Results

Define the ε-subgradient [34, Sec. 3.3] ($\varepsilon > 0$):

$$\partial_\varepsilon r(\boldsymbol{x}) \triangleq \left\{\boldsymbol{g} \in \mathbb{R}^p \,|\, r(\boldsymbol{z}) \geq r(\boldsymbol{x}) + (\boldsymbol{z}-\boldsymbol{x})^T\boldsymbol{g} - \varepsilon, \forall \boldsymbol{z} \in \mathbb{R}^p\right\} \tag{7}$$

and an *inexact proximal operator* [14]:

Definition 1: We say that $\boldsymbol{x}$ approximates $\operatorname{prox}_{ur}(\boldsymbol{a})$ with ε-precision, denoted

$$\boldsymbol{x} \approxeq_\varepsilon \operatorname{prox}_{ur}\boldsymbol{a} \tag{8}$$

if $(\boldsymbol{a}-\boldsymbol{x})/u \in \partial_{\frac{\varepsilon^2}{2u}} r(\boldsymbol{x})$.

Proposition 1: $\boldsymbol{x} \approxeq_\varepsilon \operatorname{prox}_{ur}\boldsymbol{a}$ implies $\|\boldsymbol{x} - \operatorname{prox}_{ur}\boldsymbol{a}\|_2 \leq \varepsilon$.

Proof: By Definition 1, the following holds for any $\boldsymbol{z}$:

$$ur(\boldsymbol{z}) \geq ur(\boldsymbol{x}) + (\boldsymbol{z}-\boldsymbol{x})^T(\boldsymbol{a}-\boldsymbol{x}) - 0.5\varepsilon^2 \tag{9a}$$

which is equivalent to

$$0.5\|\boldsymbol{z}-\boldsymbol{a}\|_2^2 + ur(\boldsymbol{z}) \geq 0.5\|\boldsymbol{x}-\boldsymbol{a}\|_2^2 + ur(\boldsymbol{x}) + 0.5\|\boldsymbol{z}-\boldsymbol{x}\|_2^2 - 0.5\varepsilon^2. \tag{9b}$$

Since $\boldsymbol{z} = \operatorname{prox}_{ur}\boldsymbol{a}$ minimizes the left-hand side of (9b), substituting it into (9b) completes the proof. ∎

Now, we adapt the results in [24, Sec. IV-A] and [23, Sec. 5.2.4] to complex Ψ using the fact that, for complex $\boldsymbol{y}$ and $\boldsymbol{p}$, $\|\boldsymbol{y}\|_1 = \max_{\|\boldsymbol{p}\|_\infty \leq 1} \operatorname{Re}(\boldsymbol{p}^H\boldsymbol{y})$. The proximal operator (5) with $\rho(\boldsymbol{x})$ in (4) can be rewritten as

$$\operatorname{prox}_{\lambda r}\boldsymbol{a} = \widehat{\boldsymbol{x}}(\widehat{\boldsymbol{p}}) \tag{10a}$$

where $\widehat{\boldsymbol{p}} \in \mathbb{C}^{p'}$ solves the *dual problem* [23], [24]:

$$\widehat{\boldsymbol{p}} = \arg\min_{\boldsymbol{p}\in H} \tfrac{1}{2}\|\mathcal{S}(\boldsymbol{p})\|_2^2 - \tfrac{1}{2}\|\mathcal{S}(\boldsymbol{p}) - \widehat{\boldsymbol{x}}(\boldsymbol{p})\|_2^2 \tag{10b}$$

and

$$H \triangleq \left\{\boldsymbol{w} \in \mathbb{C}^{p'} \mid \|\boldsymbol{w}\|_\infty \leq 1\right\} \tag{10c}$$

$$\mathcal{S}(\boldsymbol{p}) \triangleq \boldsymbol{a} - \lambda\operatorname{Re}(\Psi\boldsymbol{p}) \tag{10d}$$

$$\widehat{\boldsymbol{x}}(\boldsymbol{p}) \triangleq P_C\big(\mathcal{S}(\boldsymbol{p})\big) \in \mathbb{R}^p. \tag{10e}$$

When $\boldsymbol{p} \in H$, the objective function in (10b) is differentiable with respect to the real and imaginary parts of $\boldsymbol{p}$. When Ψ is real-valued, the optimal $\widehat{\boldsymbol{p}}$ must be real-valued and hence (10b) reduces to optimization with respect to $\boldsymbol{p}$ over the unit hypercube.

The duality gap for the optimization problem (10b) is

$$G(\boldsymbol{p}) = \lambda\left\{\rho\big(\widehat{\boldsymbol{x}}(\boldsymbol{p})\big) - \widehat{\boldsymbol{x}}^T(\boldsymbol{p})\operatorname{Re}(\Psi\boldsymbol{p})\right\} + I_H(\boldsymbol{p}). \tag{11}$$

To simplify the notation, we omit the dependence of $G(\boldsymbol{p})$, $\widehat{\boldsymbol{x}}(\boldsymbol{p})$, $\mathcal{S}(\boldsymbol{p})$ and $\widehat{\boldsymbol{p}}$ on $\boldsymbol{a}$ and λ. We will add the subscripts "$_{\boldsymbol{a},\lambda}$" to these quantities when we wish to emphasize their dependence on $\boldsymbol{a}$ and λ.

The following proposition extends the result in [14, Sec. 2.1] to accommodate the composite penalty (3b) that includes the indicator function $I_C(\boldsymbol{x})$; if $C = \mathbb{R}^p$, it reduces to [14, Prop. 2.3]. It can be used to guarantee the ε-precision of the proximal mapping in (8).

Proposition 2: If the duality gap (11) satisfies $G(\boldsymbol{p}) \leq \varepsilon^2/2$, then

$$\widehat{\boldsymbol{x}}(\boldsymbol{p}) \approxeq_\varepsilon \operatorname{prox}_{\lambda r}\boldsymbol{a}. \tag{12}$$

Proof: Finite $G(\boldsymbol{p})$ implies $\boldsymbol{p} \in H$. Therefore, $0 \geq \boldsymbol{z}^T\operatorname{Re}(\Psi\boldsymbol{p}) - \rho(\boldsymbol{z})$ for all $\boldsymbol{z} \in \mathbb{R}^p$ (see also (4)) and thus

$$G(\boldsymbol{p})/\lambda \geq \rho\big(\widehat{\boldsymbol{x}}(\boldsymbol{p})\big) + [\boldsymbol{z} - \widehat{\boldsymbol{x}}(\boldsymbol{p})]^T\operatorname{Re}(\Psi\boldsymbol{p}) - \rho(\boldsymbol{z}). \tag{13}$$

Use the projection theorem [34, Prop. 1.1.9 in App. B] to obtain

$$I_C(\boldsymbol{z}) \geq [\boldsymbol{z} - \widehat{\boldsymbol{x}}(\boldsymbol{p})]^T[\mathcal{S}(\boldsymbol{p}) - \widehat{\boldsymbol{x}}(\boldsymbol{p})]/\lambda. \tag{14}$$

Adding (13), (14), and $\varepsilon^2/(2\lambda) \geq G_\lambda(\boldsymbol{p})/\lambda$ and reorganizing yields

$$r(\boldsymbol{z}) \geq r\big(\widehat{\boldsymbol{x}}(\boldsymbol{p})\big) + [\boldsymbol{z} - \widehat{\boldsymbol{x}}(\boldsymbol{p})]^T[\boldsymbol{a} - \widehat{\boldsymbol{x}}(\boldsymbol{p})]/\lambda - \varepsilon^2/(2\lambda) \tag{15}$$

where we used (10d), (3b), and $I_C\big(\widehat{\boldsymbol{x}}(\boldsymbol{p})\big) = 0$. According to Definition 1, (12) and (15) are equivalent.

We introduce representative NLL functions (Section II), describe the proposed PNPG signal reconstruction algorithm (Section III), establish its convergence properties (Section IV), present numerical examples (Section V), and make concluding remarks (Section VI).

II. Probabilistic Measurement Models

For numerical stability, we normalize the likelihood function so that the corresponding NLL $\mathcal{L}(\boldsymbol{x})$ is lower-bounded by zero.

A. Poisson Generalized Linear Model

Generalized linear models (GLMs) with Poisson observations are often adopted in astronomic, optical, hyperspectral, and tomographic imaging [3], [4], [35] and are used to model event counts, e.g., numbers of particles hitting a detector. Assume that the measurements $\boldsymbol{y} = (y_n)_{n=1}^N \in \mathbb{N}_0^N$ are independent Poisson-distributed[1] with means $[\boldsymbol{\phi}(\boldsymbol{x})]_n$.

Upon normalization, we obtain the generalized Kullback-Leibler divergence form of the NLL [36]

$$\mathcal{L}(\boldsymbol{x}) = \mathbf{1}^T[\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{y}] + \sum_{n,\,y_n\neq 0} y_n \ln\frac{y_n}{[\boldsymbol{\phi}(\boldsymbol{x})]_n}. \tag{16a}$$

The NLL $\mathcal{L}(\boldsymbol{x}) : \mathbb{R}^p \mapsto \mathbb{R}_+$ is a convex function of the signal $\boldsymbol{x}$. Here, the relationship between the linear predictor $\Phi\boldsymbol{x}$ and the expected value $\boldsymbol{\phi}(\boldsymbol{x})$ of the measurements $\boldsymbol{y}$ is summarized by the link function $\boldsymbol{g}(\cdot) : \mathbb{R}^N \mapsto \mathbb{R}^N$ [37]:

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{g}^{-1}(\Phi\boldsymbol{x}). \tag{16b}$$

Note that $\operatorname{cl}(\operatorname{dom}\mathcal{L}) = \{\boldsymbol{x} \in \mathbb{R}^{\mathrm{p}} \mid \boldsymbol{\phi}(\boldsymbol{x}) \succeq \mathbf{0}\}$.

Two typical link functions in the Poisson GLM are log (described in [38, Section I-A2], see also [39]) and identity:

$$\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{\mu} - \boldsymbol{b}, \;\; \boldsymbol{\phi}(\boldsymbol{x}) = \Phi\boldsymbol{x} + \boldsymbol{b} \tag{17}$$

[1]Here, we use the extended Poisson probability mass function (pmf) $\text{Poisson}(y \,|\, \mu) = (\mu^y/y!)\,\mathrm{e}^{-\mu}$ for all $\mu \geq 0$ by defining $0^0 = 1$ to accommodate the identity-link model.

used for modeling the photon count in optical imaging and radiation activity in emission tomography [3, Ch. 9.2], as well as for astronomical image deconvolution. Here, $\Phi \in \mathbb{R}_+^{N\times p}$ and $\boldsymbol{b} \in \mathbb{R}_+^{N\times 1}$ are the known sensing matrix and intercept term, respectively; the intercept $\boldsymbol{b}$ models background radiation and scattering estimated, e.g., by calibration before the measurements $\boldsymbol{y}$ have been collected. The nonnegative set C in (2) satisfies (1), where we have used the fact that the elements of Φ are nonnegative. If $\boldsymbol{b}$ has zero components, $C \setminus \operatorname{dom} \mathcal{L}$ is *not empty* and the NLL does not have a Lipschitz-continuous gradient.

B. Linear Model With Gaussian Noise

The linear measurement model with zero-mean additive white Gaussian noise (AWGN) leads to the following scaled NLL:

$$\mathcal{L}(\boldsymbol{x}) = \tfrac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 \tag{18}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ is the measurement vector, and constant terms (not functions of $\boldsymbol{x}$) have been ignored. This NLL belongs to the Gaussian GLM with identity link without intercept: $\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{\mu}$. Here, $\operatorname{dom} \mathcal{L}(\boldsymbol{x}) = \mathbb{R}^{\mathrm{p}}$, any closed convex C satisfies (1), and the set $C \setminus \operatorname{dom} \mathcal{L}$ is empty.

Minimization of the objective function (3a) with Gaussian NLL (18) and penalty (3b) with $\rho(\boldsymbol{x})$ in (4) is an *analysis basis pursuit denoising (BPDN) problem with a convex signal constraint*.

III. Reconstruction Algorithm

We propose a PNPG approach for minimizing (3a) that combines convex-set projection with Nesterov acceleration [15], [16] and applies adaptive step size to adapt to the local curvature of the NLL and restart to ensure monotonicity of the resulting iteration. The pseudo code in Algorithm 1 summarizes our PNPG method.

Define the quadratic approximation of the NLL $\mathcal{L}(\boldsymbol{x})$ as

$$Q_\beta(\boldsymbol{x} \mid \bar{\boldsymbol{x}}) = \mathcal{L}(\bar{\boldsymbol{x}}) + (\boldsymbol{x} - \bar{\boldsymbol{x}})^T \nabla\mathcal{L}(\bar{\boldsymbol{x}}) + \frac{1}{2\beta}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2 \tag{19}$$

with step-size tuning constant $\beta > 0$. Iteration i of the PNPG method proceeds as follows:

$$B^{(i)} = \beta^{(i-1)}/\beta^{(i)} \tag{20a}$$

$$\theta^{(i)} = \begin{cases} 1, & i \le 1 \\ \dfrac{1}{\gamma} + \sqrt{b + B^{(i)}(\theta^{(i-1)})^2}, & i > 1 \end{cases} \tag{20b}$$

$$\Theta^{(i)} = (\theta^{(i-1)} - 1)/\theta^{(i)} \tag{20c}$$

$$\bar{\boldsymbol{x}}^{(i)} = P_C\big(\boldsymbol{x}^{(i-1)} + \Theta^{(i)}(\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^{(i-2)})\big) \tag{20d}$$

$$\boldsymbol{x}^{(i)} = \operatorname{prox}_{\beta^{(i)}ur}\big(\bar{\boldsymbol{x}}^{(i)} - \beta^{(i)}\nabla\mathcal{L}(\bar{\boldsymbol{x}}^{(i)})\big) \tag{20e}$$

where $\beta^{(i)} > 0$ is an *adaptive step size* chosen to satisfy the *majorization condition*

$$\mathcal{L}(\boldsymbol{x}^{(i)}) \le Q_{\beta^{(i)}}(\boldsymbol{x}^{(i)} \mid \bar{\boldsymbol{x}}^{(i)}) \tag{21}$$

Algorithm 1: PNPG iteration.

Input: $\boldsymbol{x}^{(0)}$, u, γ, b, $\mathtt{m}$, $\mathtt{m}$, ξ, η, and threshold ϵ
Output: $\arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$
Initialization: $\boldsymbol{x}^{(-1)} \leftarrow \boldsymbol{0}$, $i \leftarrow 0$, $\kappa \leftarrow 0$, $\beta^{(1)}$ by the BB method
repeat
  $i \leftarrow i + 1$ and $\kappa \leftarrow \kappa + 1$
  while *true* **do** `// backtracking search`
    evaluate (20a) to (20d)
    if $\bar{\boldsymbol{x}}^{(i)} \notin \operatorname{dom}\mathcal{L}$ **then** `// domain restart`
      $\theta^{(i-1)} \leftarrow 1$ and continue
    solve the proximal mapping in (20e)
    if *majorization condition* (21) *holds or the number of backtrackings exceeds* t_{MAX} **then**
      break
    else
      if $\beta^{(i)} > \beta^{(i-1)}$ **then** `// increase` $\mathtt{m}$
        $\mathtt{m} \leftarrow \mathtt{m} + \mathtt{m}$
      $\beta^{(i)} \leftarrow \xi\beta^{(i)}$ and $\kappa \leftarrow 0$
  if $f(\boldsymbol{x}^{(i)}) > f(\boldsymbol{x}^{(i-1)})$ **then**
    if (21) *holds* **then**
      if $\bar{\boldsymbol{x}}^{(i)} \neq \boldsymbol{x}^{(i-1)}$ *and* $f(\boldsymbol{x}^{(i)}) \le f(\bar{\boldsymbol{x}}^{(i)})$ **then** `// function restart`
        $\theta^{(i-1)} \leftarrow 1$, $i \leftarrow i - 1$, and continue
      if $\eta > \eta_{\mathrm{MIN}}$ *and* $f(\boldsymbol{x}^{(i)}) > f(\bar{\boldsymbol{x}}^{(i)})$ **then** `// more accurate proximal`
        $\eta \leftarrow \eta/10$, $i \leftarrow i - 1$, and continue
    declare convergence
  if *convergence cond.* (23a) *holds with threshold* ϵ **then**
    declare convergence
  if $\kappa \ge \mathtt{m}$ **then** `// adapt step size`
    $\kappa \leftarrow 0$ and $\beta^{(i+1)} \leftarrow \beta^{(i)}/\xi$
  else
    $\beta^{(i+1)} \leftarrow \beta^{(i)}$
until *convergence declared or maximum number of iterations exceeded*

using a simple adaptation scheme that aims at keeping $\beta^{(i)}$ as large as possible; see also Section III-B and Algorithm 1. Here,

$$\gamma \ge 2, \qquad b \in [0, 1/4] \tag{22}$$

in (20b) are *momentum tuning constants*. We will denote $\theta^{(i)}$ as $\theta^{(i)}_{\gamma,b}$ when we wish to emphasize its dependence on γ and b. We declare convergence when

$$\sqrt{\delta^{(i)}} \le \epsilon\|\boldsymbol{x}^{(i)}\|_2 \tag{23a}$$

where $\epsilon > 0$ is the convergence threshold and $\delta^{(i)}$ is the local variation of signal iterates:

$$\delta^{(i)} \triangleq \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i-1)}\|_2^2. \tag{23b}$$

We need $B^{(i)}$ in (20a) to derive the theoretical guarantee for the convergence speed of the PNPG iteration and its sequence convergence. A similar idea for handling the increasing step size

in its TFOCS framework is seen in [18]. However, [18] does not address this modification in detail or establish convergence properties of the corresponding method.

The acceleration step (20d) extrapolates the two latest iteration points in the direction of their difference $\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^{(i-2)}$, followed by the projection onto the convex set C, which has also been proposed in our preliminary work [8] and in the variable-metric/scaling method [10]. For nonnegative C in (2), this projection has closed form; see (6c). If C is an intersection of convex sets with a simple individual projection operator for each, we can apply projections onto convex sets (POCS) [6].

For $\rho(\boldsymbol{x})$ in (4), we compute the proximal mapping (20e) using the dual formulation in (10) and a simpler version of PNPG, Nesterov's projected-gradient algorithm, because the proximal step in this case reduces to projection onto H in (10c); for the TV penalty, this method is similar to the TV denoising scheme in [24]. Because of its iterative nature, (20e) is *inexact*; this inexactness can be modeled as

$$\boldsymbol{x}^{(i)} \approx_{\varepsilon^{(i)}} \operatorname{prox}_{\beta^{(i)}ur}\left(\bar{\boldsymbol{x}}^{(i)} - \beta^{(i)}\nabla\mathcal{L}\left(\bar{\boldsymbol{x}}^{(i)}\right)\right) \tag{24}$$

where $\varepsilon^{(i)}$ quantifies the precision of the PG step in Iteration i.

If we remove the convex-set constraint by setting $C = \mathbb{R}^p$, iteration (20a)–(20e) reduces to the Nesterov's proximal-gradient iteration with adaptive step size that imposes signal sparsity *only* in the analysis form (termed NPG_S); see Section V-B for an illustrative comparison between NPGs and PNPG.

We now extend [16, Lemma 2.3] to the inexact proximal operation:

Lemma 1: Assume convex and differentiable NLL $\mathcal{L}(\boldsymbol{x})$ and convex $\rho(\boldsymbol{x})$, and consider an inexact PG step (24) with step size $\beta^{(i)}$ that satisfies the majorization condition (21). Then,

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^{(i)}) \geq \frac{1}{2\beta^{(i)}}\left[\|\boldsymbol{x}^{(i)} - \boldsymbol{x}\|_2^2 - \|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{x}\|_2^2 - (\varepsilon^{(i)})^2\right] \tag{25}$$

for all $i \geq 1$ and any $\boldsymbol{x} \in \mathbb{R}^p$.

Proof: See Appendix A. ■

Lemma 1 is general and algorithm independent, because $\bar{\boldsymbol{x}}^{(i)}$ can be any value in dom $\mathcal{L}$ and we have used only the fact that step size $\beta^{(i)}$ satisfies the majorization condition (21), rather than depending on specific details of the step-size selection. We use this result to establish the monotonicity property in Proposition 3 and to derive and analyze our accelerated PG scheme.

A. Restart and Monotonicity

If $f(\bar{\boldsymbol{x}}^{(i)}) > f(\boldsymbol{x}^{(i)}) > f(\boldsymbol{x}^{(i-1)})$ or $\bar{\boldsymbol{x}}^{(i)} \in C \setminus \text{dom}\,\mathcal{L}$, set

$$\theta^{(i-1)} = 1 \qquad \text{(restart)}, \tag{26}$$

re-evaluate (20b)–(20e), and refer to this action as *function restart* [40] or *domain restart*, respectively; see Algorithm 1. Function and domain restarts ensure that the PNPG iteration is monotonic and $\bar{\boldsymbol{x}}^{(i)}$ remains within dom f as long as the projected initial value is within dom f: $f(P_C(\boldsymbol{x}^{(0)})) < +\infty$. In this paper, we employ PNPG iteration *with* restart, unless specified otherwise (e.g., in Theorems 1 and 2 in Section IV).

Proposition 3 (Monotonicity): The inexact PG step (24) is monotonic:

$$f(\boldsymbol{x}^{(i)}) \leq f\left(\bar{\boldsymbol{x}}^{(i)}\right) \tag{27a}$$

if it is sufficiently accurate such that

$$\varepsilon^{(i)} \leq \|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}\|_2. \tag{27b}$$

Hence, the PNPG iteration with restart and inexact PG steps (24) is non-increasing:

$$f(\boldsymbol{x}^{(i)}) \leq f\left(\boldsymbol{x}^{(i-1)}\right) \tag{28}$$

if (27b) holds for all i.

Proof: (27a) is straightforward by plugging $\boldsymbol{x} = \bar{\boldsymbol{x}}^{(i)}$ and (27b) into (25).

If there is no restart in Iteration i, the objective function has not increased. If there is a restart, $\theta^{(i-1)} = 1$, (20d) simplifies to $\bar{\boldsymbol{x}}^{(i)} = P_C(\boldsymbol{x}^{(i-1)}) = \boldsymbol{x}^{(i-1)}$, and monotonicity follows due to $\bar{\boldsymbol{x}}^{(i)} = \boldsymbol{x}^{(i-1)}$. ■

To establish the monotonicity in Proposition 3, the step size $\beta^{(i)}$ need satisfy only the majorization condition (21).

B. Adaptive Step Size

Define the *step-size adaptation parameter*

$$\xi \in (0, 1). \tag{29}$$

We propose the following adaptive scheme for selecting $\beta^{(i)}$:

i)
- if there have been no step-size backtracking events or increase attempts for $\mathtt{m}$ consecutive iterations ($i-\mathtt{m}$ to $i-1$), start with a larger step size:

$$\bar{\beta}^{(i)} = \beta^{(i-1)}/\xi \qquad \text{(increase attempt)}; \tag{30a}$$

- otherwise start with

$$\bar{\beta}^{(i)} = \beta^{(i-1)}; \tag{30b}$$

ii) (backtracking search) select

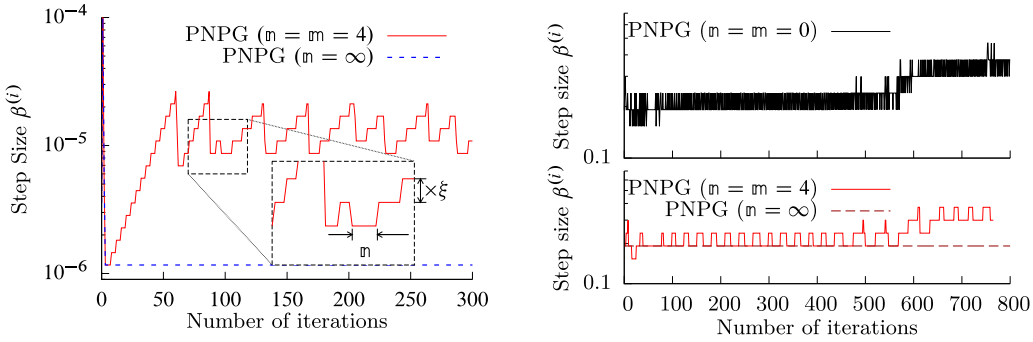
$$\beta^{(i)} = \xi^{t_i}\bar{\beta}^{(i)} \tag{30c}$$

where $0 \leq t_i \leq t_{\text{MAX}}$ is the smallest integer such that (30c) satisfies (21); *backtracking event* corresponds to $t_i > 0$.

iii) if $\max(\beta^{(i)}, \beta^{(i-1)}) < \bar{\beta}^{(i)}$, increase $\mathtt{m}$ by a nonnegative integer $\mathtt{m}$:

$$\mathtt{m} \leftarrow \mathtt{m} + \mathtt{m} \tag{30d}$$

We select the initial step size $\bar{\beta}^{(1)}$ using the BB method [41]. If there has been an attempt to change the step size in any of the previous $\mathtt{m}$ consecutive iterations, we start the backtracking search ii) with the step size from the latest completed iteration. Consequently, $\beta^{(i)}$ will be approximately piecewise constant as a function of the iteration index i; see Fig. 1, which shows the evolutions of $\beta^{(i)}$ for measurements following the Poisson generalized linear and Gaussian linear models corresponding to Figs. 4(a) and 6(b) in Sections V-A and V-B. To reduce



(a) Poisson model with identity link.

(b) Gaussian linear model.

Fig. 1. Step sizes $\beta^{(i)}$ as functions of the number of iterations for Poisson and Gaussian linear models.

sensitivity to the choice of the tuning constant $\mathbb{m}$, we increase its value by $\bar{\mathbb{m}}$ if there is a failed attempt to increase the step size in Iteration i; i.e., $\bar{\beta}^{(i)} > \beta^{(i-1)}$ and $\beta^{(i)} < \bar{\beta}^{(i)}$.

The adaptive step-size strategy keeps $\beta^{(i)}$ as large as possible subject to (21), which is important not only because the signal iterate may reach regions of $\mathcal{L}(\boldsymbol{x})$ with different local Lipschitz constants, but also because of the varying curvature of $\mathcal{L}(\boldsymbol{x})$ in different updating directions. For example, a (backtracking-only) PG-type algorithm with non-adaptive step size would fail or converge very slowly if the local Lipschitz constant of $\nabla\mathcal{L}(\boldsymbol{x})$ decreases as the algorithm iterates, because the step size will not adjust and track this decrease; see also Section V, which demonstrates the benefits of step-size adaptation.

Setting $\mathbb{m} = +\infty$ corresponds to step-size backtracking only. A step-size adaptation scheme with $\mathbb{m} = \bar{\mathbb{m}} = 0$ initializes the step-size search aggressively, with an increase attempt (30a) in each iteration.

C. Inner-Iteration Warm Start and Convergence Criteria

For $\rho(\boldsymbol{x})$ in (4), the inner iteration solves (20e) using the dual problem (10b). Denote by $\boldsymbol{p}^{(i,j)}$ the iterates of the dual variable $\boldsymbol{p}$ in the jth inner iteration step within Iteration i; this inner iteration solves (20e) using (10b). The initial $\boldsymbol{p}^{(i,0)}$ is the latest $\boldsymbol{p}$ from Iteration $i-1$, which is referred to in [14] as the *warm restart*. (The variable metric inexact line-search algorithm (VMILA) [19] also uses warm restart.)

We consider two convergence criteria. The first tracks local variation of the signal iterates (23b):

$$\|\boldsymbol{x}^{(i,j)} - \boldsymbol{x}^{(i,j-1)}\| \leq \eta\sqrt{\delta^{(i-1)}} \tag{31a}$$

where η is a tuning constant.

The second *duality-gap–based* criterion relies on the result in Proposition 2 to guarantee that $(\theta^{(k)}\varepsilon^{(k)})^2$ decreases at a rate of $\mathcal{O}(k^{-q})$ within each iteration segment without restart; this guarantee allows us to control the decrease of the convergence-rate upper bound in Section IV. Denote by ι_i the iteration index of the latest restart prior to (and excluding) Iteration i ($i \geq 1$); set its initial value $\iota_1 = 0$. We select the duality-gap based inner-iteration convergence criterion as (see also (10e) and (11))

$$\frac{G^{(i,j)}}{\beta^{(i)}u\,\rho\big(\widehat{\boldsymbol{x}}^{(i,j)}\big)} \leq \frac{\eta}{(i-\iota_i)^q(\theta^{(i)})^2} \tag{31b}$$

where η is a tuning constant and q is the *accuracy rate* [14]. Here, $G^{(i,j)}$ and $\widehat{\boldsymbol{x}}^{(i,j)}$ are the duality gap $G_{\boldsymbol{a},\lambda}(\boldsymbol{p}^{(i,j)})$ and $\widehat{\boldsymbol{x}}_{\boldsymbol{a},\lambda}(\boldsymbol{p}^{(i,j)})$ in (11) and (10e) (respectively) for $\boldsymbol{a} = \bar{\boldsymbol{x}}^{(i)} - \beta^{(i)}\nabla\mathcal{L}(\bar{\boldsymbol{x}}^{(i)})$ and $\lambda = \beta^{(i)}u$. Without restart (i.e., $\iota_i \equiv 0$) and step-size adaptation, (31b) reduces to the inner-iteration convergence criterion in [14, Sec. 6.1].

Adjusting η. We use η in (31a)–(31b) to trade off accuracy and speed of the inner iterations. If $f(\boldsymbol{x}^{(i)}) > \max\big(f(\boldsymbol{x}^{(i-1)}), f(\bar{\boldsymbol{x}}^{(i)})\big)$ indicating that the monotonicity condition (27b) does not hold, we decrease η by an order of magnitude (10 times) and re-evaluate (20a)–(20e).

IV. Convergence Analysis

We now bound the convergence rate of the PNPG method without restart.

Theorem 1 (Convergence of the Objective Function): Assume that the NLL $\mathcal{L}(\boldsymbol{x})$ is convex and differentiable, $\rho(\boldsymbol{x})$ is convex, the closed convex set C satisfies

$$C \subseteq \operatorname{dom}\mathcal{L} \tag{32}$$

(implying no need for domain restart), and the momentum tuning constants are within the range (22). Consider the PNPG iteration without restart with the inexact PG step (24) in place of (20e). The convergence rate of the PNPG iteration is bounded as follows: for $k \geq 1$,

$$\Delta^{(k)} \leq \frac{\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2 + \mathcal{E}^{(k)}}{2\beta^{(k)}(\theta^{(k)})^2} \tag{33a}$$

$$\leq \gamma^2\frac{\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2 + \mathcal{E}^{(k)}}{2\big(\sqrt{\beta^{(1)}} + \sum_{i=1}^k\sqrt{\beta^{(i)}}\big)^2} \tag{33b}$$

where $\boldsymbol{x}^\star$ is a minimum point of $f(\boldsymbol{x})$ and

$$\Delta^{(k)} \triangleq f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star) \tag{34a}$$

$$\mathcal{E}^{(k)} \triangleq \sum_{i=1}^k(\theta^{(i)}\varepsilon^{(i)})^2 \tag{34b}$$

are the centered objective function and the cumulative error term, which accounts for the inexact PG steps, respectively.

Proof: See Appendix A for the proof of (33a); then, to obtain (33b), use

$$\theta^{(k)}\sqrt{\beta^{(k)}} \geq \frac{1}{\gamma}\sqrt{\beta^{(k)}} + \theta^{(k-1)}\sqrt{\beta^{(k-1)}} \tag{35a}$$

$$\geq \frac{1}{\gamma}\sum_{i=2}^{k}\sqrt{\beta^{(i)}} + \theta^{(1)}\sqrt{\beta^{(1)}} \tag{35b}$$

for all $k > 1$, where (35a) follows from the definitions of $B^{(k)}$ and $\theta^{(k)}$ in (20a) and (20b), and (35b) follows by repeated application of the inequality (35a) with k replaced by $k-1, k-2, \ldots, 2$. ■

Theorem 1 shows that better initialization, smaller proximal-mapping approximation error, and larger step sizes $(\beta^{(i)})_{i=1}^{k}$ help lower the convergence-rate upper bounds in (33). This result motivates our step-size adaptation with the goal of maintaining large $(\beta^{(i)})_{i=1}^{k}$; see Section III-B. To derive this theorem, we have used only the fact that the step size $\beta^{(i)}$ satisfies the majorization condition (21), rather than taking advantage of specific details of the step-size selection.

To minimize the upper bound in (33a), we can select $\theta^{(i)}$ to satisfy (A16b) with equality, which corresponds to $\theta^{(i)}_{2,1/4}$ in (20b), on the boundary of the feasible region in (22). By (35a), $\sqrt{\beta^{(k)}}\theta^{(k)}$ and the denominator of the bound in (33a) are strictly increasing sequences. The upper bound in (33b) is not a function of b and is minimized with respect to γ for $\gamma = 2$, given the fixed step sizes $(\beta^{(i)})_{i=0}^{+\infty}$.

Corollary 1: Under the assumptions of Theorem 1, the convergence of PNPG iteration $\boldsymbol{x}^{(k)}$ without restart is bounded as follows:

$$\Delta^{(k)} \leq \gamma^2\frac{\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^{\star}\|_2^2 + \mathcal{E}^{(k)}}{2(k+1)^2\beta_{\min}} \tag{36a}$$

for $k \geq 1$, provided that

$$\beta_{\min} \triangleq \min_{k=1}^{+\infty}\beta^{(k)} > 0. \tag{36b}$$

Proof: Use (33b) and the fact that $\sqrt{\beta^{(1)}} + \sum_{i=1}^{k}\sqrt{\beta^{(i)}} \geq (k+1)\sqrt{\beta_{\min}}$. ■

The assumption (36b) is implied by, and weaker than, the Lipschitz continuity of $\nabla\mathcal{L}(\boldsymbol{x})$; indeed, $\beta_{\min} > \xi/L$ if $\nabla\mathcal{L}(\boldsymbol{x})$ has a Lipschitz constant L; see also (29).

According to Corollary 1, the PNPG iteration attains the $\mathcal{O}(k^{-2})$ convergence rate as long as the step size $\beta^{(i)}$ is bounded away from zero (see (36b)) and the cumulative error term (34b) converges:

$$\mathcal{E}^{(+\infty)} \triangleq \lim_{k\to+\infty}\mathcal{E}^{(k)} < +\infty \tag{37}$$

which requires that $(\theta^{(k)}\varepsilon^{(k)})^2$ decreases at a rate of $\mathcal{O}(k^{-q})$ with $q > 1$. This condition, also key for establishing convergence of iterates in Theorem 2, motivates us to use decreasing convergence criteria (31a)–(31b) for the inner proximal-mapping iterations, where (31b) guarantees (37) upon choosing an appropriate q.

We contrast our result in Theorem 1 with existing work on accommodating inexact proximal mappings in PG schemes. By recursively generating a function sequence that approximates the objective function, [14] gives an asymptotic analysis of the effect of $\varepsilon^{(i)}$ on the convergence rate of accelerated PG methods with inexact proximal mapping. However, no explicit upper bound is provided for $\Delta^{(k)}$. Schmidt *et al.* [13] provide convergence-rate analysis and an upper bound on $\Delta^{(k)}$, but their analysis does not apply here because it relies on fixed step-size assumption, uses different form of acceleration [13, Prop. 2], and has no convex-set constraint. Bonettini *et al.* [19] analyze the inexactness of proximal mapping but for proximal variable-metric/scaling methods with $\mathcal{O}(k^{-1})$ convergence rate for the objective function.

We now establish convergence of the PNPG iterates.

Theorem 2 (Convergence of Iterates): Assume that

1) the conditions of Theorem 1 hold;
2) $\mathcal{E}^{(+\infty)}$ exists: (37) holds;
3) the momentum tuning constants (γ, b) satisfy

$$\gamma > 2, \quad b \in [0, 1/\gamma^2]; \tag{38}$$

4) the step-size sequence $(\beta^{(i)})_{i=1}^{+\infty}$ is bounded within the range $[\beta_{\min}, \beta_{\max}]$, for $\beta_{\min} > 0$.

Consider the PNPG iteration without restart with the inexact PG step (24) in place of (20e). Then, the sequence of PNPG iterates $\boldsymbol{x}^{(i)}$ converges to a minimizer of $f(\boldsymbol{x})$.

Proof: See Appendix B. ■

Observe that Assumption 3 requires a narrower range of (γ, b) than (22): indeed (38) is a strict subset of (22). The intuition is to leave a sufficient gap between the two sides of (A16a) so that their difference becomes a quantity that is roughly proportional to the growth of $\theta^{(i)}$, which is important for proving the convergence of signal iterates [42]. Although the momentum term (20b) with $\gamma = 2$ is optimal in terms of minimizing the upper bound on the convergence rate (see Theorem 1), it appears difficult or impossible to prove convergence of the signal iterates $\boldsymbol{x}^{(i)}$ for this choice of γ because, in this case, the gap between the two sides of (A16a) is upper-bounded by a constant.

Iterate convergence results in [10], [42], [43] apply to momentum-accelerated methods that require non-increasing step-size sequences and do not adjust to the local curvature of the NLL. Aujol and Dossal [43] analyze both the convergence of the objective function and the iterates with inexact proximal operator for $B^{(1)} = 1$ and $\mathbb{n} = \infty$, i.e., with decreasing step size only, and for a different (less aggressive) $\theta^{(i)}$ than ours in (20b). Bonettini *et al.* use the ideas from [42] to establish convergence of iterates for their variable-metric/scaling approach in [10], but this analysis does not account for inexact proximal steps.

A. $\mathcal{O}(k^{-2})$ Convergence Acceleration Approaches

A few variants that accelerate the PG method achieve the $\mathcal{O}(k^{-2})$ convergence rate [18, Sec. 5.2]. One competitor proposed by Auslender and Teboulle in [20, Sec. 5] and restated

in [18] where it was referred to as AT, replaces (20d)–(20e) with

$$\bar{\boldsymbol{x}}^{(i)} = \Big(1 - \frac{1}{\theta^{(i)}}\Big)\boldsymbol{x}^{(i-1)} + \frac{1}{\theta^{(i)}}\widetilde{\boldsymbol{x}}^{(i-1)} \tag{39a}$$

$$\widetilde{\boldsymbol{x}}^{(i)} = \operatorname{prox}_{\theta^{(i)}\beta^{(i)}ur}\big(\widetilde{\boldsymbol{x}}^{(i-1)} - \theta^{(i)}\beta^{(i)}\nabla\mathcal{L}(\bar{\boldsymbol{x}}^{(i)})\big) \tag{39b}$$

$$\boldsymbol{x}^{(i)} = \Big(1 - \frac{1}{\theta^{(i)}}\Big)\boldsymbol{x}^{(i-1)} + \frac{1}{\theta^{(i)}}\widetilde{\boldsymbol{x}}^{(i)} \tag{39c}$$

where $\theta^{(i)} = \theta^{(i)}_{2,1/4}$ in (20b). Here, $\beta^{(i)}$ in the TFOCS implementation [18] is selected using the aggressive search with $\mathbb{n} = \mathbb{m} = 0$.

All intermediate signals in (39a)–(39c) belong to C and do not require projections onto C. However, as $\theta^{(i)}$ increases with i, step (39b) becomes unstable, especially when an iterative solver is needed for its proximal operation. To stabilize its convergence, AT relies on periodic restart by resetting $\theta^{(i)}$ using (26) [18]. However, the period of restart is a tuning parameter that is not easy to select. For a linear Gaussian model, this period varies according to the condition number of the sensing matrix Φ [18], which is generally unavailable and not easy to compute for large-scale problems. For other models, there are no guidelines how to select the restart period.

In Section V, we show that AT converges slowly compared with PNPG, which justifies the use of projection onto C in (20d) and (20d)–(20e) instead of (39a)–(39c). PNPG usually runs uninterrupted (without restart) over long stretches and benefits from Nesterov's acceleration within these stretches, which may explain its better convergence properties compared with AT. PNPG may also be less sensitive than AT to proximal-step inaccuracies; we have established convergence-rate bounds for PNPG that account for inexact proximal steps (see (33) and (36a)), whereas AT does not yet have such bounds, to the best of our knowledge.

1) Relationship With FISTA: The PNPG method is a generalized FISTA [16] that accommodates convex constraints, more general NLLs,[2] and (increasing) adaptive step size. In contrast with PNPG, FISTA has a non-increasing step size $\beta^{(i)}$, which allows for setting $B^{(i)} = 1$ in (20b) for all i (see Appendix A-II); upon setting $(\gamma, b) = (2, 1/4)$, this choice yields the standard FISTA (and Nesterov's [15]) update. Convergence of signal iterates has not been established for FISTA with $(\gamma, b) = (2, 1/4)$ [44]. Theorem 2 comes close to this goal because it establishes convergence of iterates of PNPG and corresponding FISTA for (γ, b) *arbitrarily close* to $(2, 1/4)$.

The method in [10] is a variable-metric/scaling version of FISTA with projection of the extrapolation step to account for the convex constraints. [21] analyzes a version of FISTA where the (decreasing) step size is adjusted using a condition in [45] different from the majorization condition (21), and establishes objective-function convergence under the assumption that the step size is lower bounded. As FISTA, [10] and [21] *do not* adapt the step size and hence do not adjust to the local curvature of the NLL.

[2]FISTA has been developed for the linear Gaussian model in Section II-B.

V. Numerical Examples

We now evaluate our proposed algorithm by means of numerical simulations. We adopt the nonnegative C in (2) and the ℓ_1-norm sparsifying penalty in (4). The PNPG iterations with the local-variation and duality-gap inner convergence criteria (31a) and (31b) are labeled PNPG and PNPG_d, respectively.

All iterative methods that we compare use the convergence criterion (23a) with

$$\epsilon = 10^{-9} \tag{40}$$

and have the maximum number of iterations $I_{\max} = 10^4$. In the presented examples, PNPG uses momentum tuning constants $(\gamma, b) = (2, 1/4)$ and adaptive step-size parameters $\mathbb{n} = 4$ (unless specified otherwise), $\mathbb{n} = \mathbb{m}$, $\xi = 0.8$, inner-iteration convergence constants $\eta = 10^{-2}$ and $(\eta, q) = (1, 1.0001)$ for PNPG and PNPG_d (respectively), and maximum number of inner iterations $J_{\max} = 1000$. Here, PNPG_d uses $q = 1.0001$ with goal to guarantee (37).

We apply the AT method (39) implemented in the TFOCS package [18] with adaptive step size and a periodic restart every 200 iterations (tuned for its best performance) and our proximal mapping. Our inner convergence criteria (31b) cannot be implemented in the TFOCS package (i.e., it require editing its code). Hence, we select the proximal mapping that has a relative-error inner convergence criterion

$$\|\boldsymbol{p}^{(i,j)} - \boldsymbol{p}^{(i,j-1)}\|_2 \le \epsilon'\|\boldsymbol{p}^{(i,j)}\|_2, \tag{41a}$$

where $\boldsymbol{p}^{(i,j)}$ is the dual variable employed by the inner iterations. This relative-error inner convergence criterion is easy to incorporate into the TFOCS software package [18] and is already used by the SPIRAL package; see [46]. Here, we select

$$\epsilon' = 10^{-6} \tag{41b}$$

for both AT and SPIRAL and set their maximum number of inner iterations to 100.

We apply the CP approach based on [23, Sec. 7.5]:

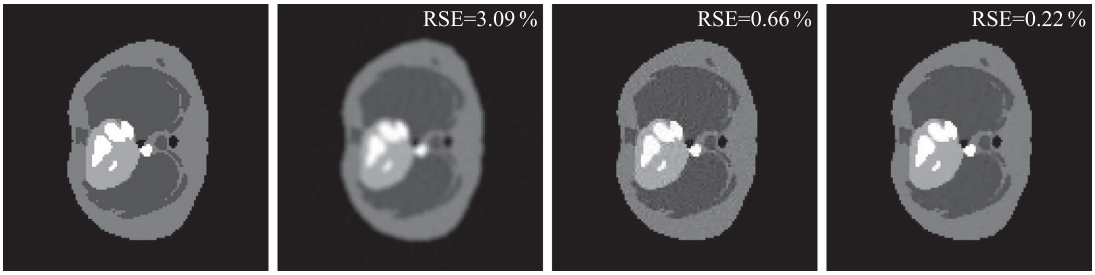
$$\boldsymbol{z} \leftarrow \operatorname{prox}_{\sigma_1 F^*}\big(\boldsymbol{z} + \sigma_1\Phi\bar{\boldsymbol{x}}\big) \tag{42a}$$

$$\boldsymbol{p} \leftarrow P_{uH}(\boldsymbol{p} + \sigma_2\Psi^H\bar{\boldsymbol{x}}) \tag{42b}$$

$$\bar{\boldsymbol{x}} \leftarrow 2P_C\Big(\boldsymbol{x} - \tau\big(\Phi^T\boldsymbol{z} + \operatorname{Re}(\Psi\boldsymbol{p})\big)\Big) - \boldsymbol{x} \tag{42c}$$

$$\boldsymbol{x} \leftarrow (\bar{\boldsymbol{x}} + \boldsymbol{x})/2 \tag{42d}$$

obtained by splitting the objective function (3a) into the sum of $F(\Phi\boldsymbol{x}) + u\|\Psi^H\boldsymbol{x}\|_1$ and $I_C(\boldsymbol{x})$, where the first summand is a convex lower semicontinuous function of $[\Phi^H\ \Psi]^H\boldsymbol{x}$ and $F(\Phi\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})$. In our examples in Sections V-A and V-B, $F(\boldsymbol{y})$ and its convex conjugate $F^*(\boldsymbol{z})$ have analytical proximal operators [23, Sec. 7.4]; hence, CP *does not* require an inner iteration. In the original CP algorithm in [27], Chambolle and Pock select $\sigma_1 = \sigma_2 = \sigma$. However, when the difference between $\|\Phi\|_2$ and $\|\Psi\|_2$ becomes large, it is hard to find tuning constants of the form $(\sigma_1, \sigma_2, \tau) = (\sigma, \sigma, \tau)$ that ensure fast convergence of the CP algorithm, which is why we do not impose $\sigma_1 = \sigma_2 = \sigma$ here. Another version of CP can be obtained by associating the regularization parameter u with Ψ instead of the ℓ_1-norm function, which leads to replacing uH with H and Ψ, Ψ^H with



(a) radio-isotope concentration (b) FBP (c) DWT (d) TV

Fig. 2. (a) True emission image and (b)–(d) the reconstructions of the emission concentration map.

$u\Psi$, $u\Psi^H$, respectively, in (42). In this paper, we adopt the version of CP in (42).

All the numerical examples were performed on a Linux workstation with an Intel Xeon CPU E31245 (3.30 GHz) and 8 GB memory. The operating system is Ubuntu 14.04 LTS (64-bit). The Matlab implementation of the proposed algorithms and numerical examples is available [47].

A. PET Image Reconstruction From Poisson Measurements

In this example, we adopt the Poisson GLM (16a) with identity link in (17). Consider PET reconstruction of the 128×128 concentration map $\boldsymbol{x}$ in Fig. 2(a), which represents simulated radiotracer activity in the human chest. Assume that the corresponding 128×128 attenuation map $\boldsymbol{\kappa}$ is known, which is needed to model the attenuation of the gamma rays [35] and compute the sensing matrix Φ in this application. We collect the photons from 90 equally spaced directions over 180°, with 128 radial samples in each direction. Here, we adopt the parallel strip-integral matrix Γ [48, Ch. 25.2] and use its implementation in the Image Reconstruction Toolbox (IRT) [49] with sensing matrix

$$\Phi = w\mathrm{diag}\big(\exp_\circ(-\Gamma\boldsymbol{\kappa} + \boldsymbol{c})\big)\Gamma \tag{43}$$

where $\boldsymbol{c}$ is a known vector generated using a zero-mean independent, identically distributed (i.i.d.) Gaussian sequence with variance 0.3 to model the detector efficiency variation; $w > 0$ is a known scaling constant controlling the expected total number of detected photons due to electron-positron annihilation; and $\mathbf{1}^T\mathrm{E}(\boldsymbol{y} - \boldsymbol{b}) = \mathbf{1}^T\Phi\boldsymbol{x}$, which is a signal-to-noise ratio (SNR) measure. Assume that the background radiation, scattering effect, and accidental coincidence combined lead to a known (generally nonzero) intercept term $\boldsymbol{b}$ in the Poisson GLM (17). The elements of the intercept term have been set to a constant equal to 10% of the sample mean of $\Phi\boldsymbol{x}_{\text{true}}$: $\boldsymbol{b} = (\mathbf{1}^T\Phi\boldsymbol{x}_{\text{true}})/(10N)\mathbf{1}$.

The above model, choices of parameters in the PET system setup, and concentration map have been adopted from IRT [49, `emission/em_test_setup.m`].

Here, we consider the DWT and isotropic TV sparsifying transforms. We use the 2-D Haar DWT with 6 decomposition levels and a full circular mask [50] to construct a sparsifying dictionary matrix $\Psi \in \mathbb{R}^{12449\times 14056}$ with orthonormal rows, i.e., $\Psi\Psi^T = I$, which allows efficient inner iteration. We compare the filtered backprojection (FBP) [35] and PG methods that aim at minimizing (3) with nonnegative C in (2) and DWT and TV sparsifying transforms.

We implemented SPIRAL with TV penalty using the centered NLL term (16a), which improves the numerical stability compared with the original code in [46]. We do not compare with SPIRAL that uses DWT penalty because its inner iteration for the proximal step requires a square orthogonal Ψ (see [4, Sec. II-B]), which is not the case here. We also compare with VMILA [19], [51] with both DWT and TV penalties and its default tuning constants, which yield good performance; hence VMILA is insensitive to tuning.

In this example, we adopt the following form of the regularization constant u:

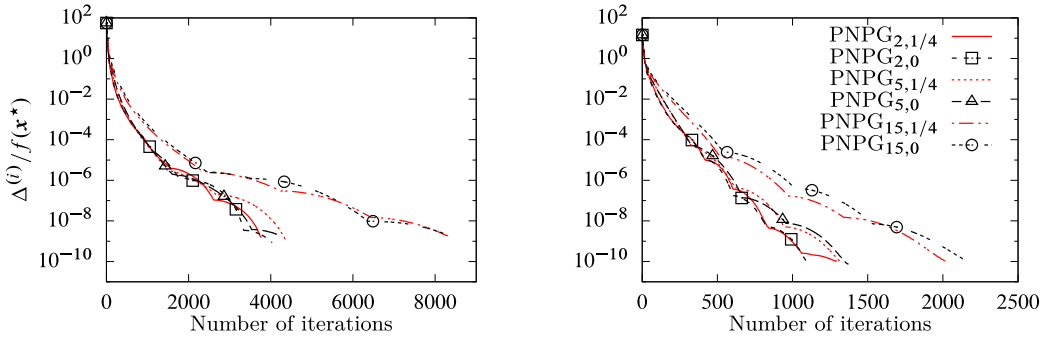
$$u = 10^a, \tag{44}$$

vary a in the range $[-6, 3]$ with a grid size of 0.5, and search for the reconstructions with the best average relative square error (RSE) performance; here, $\mathrm{RSE} = \|\widehat{\boldsymbol{x}} - \boldsymbol{x}_{\text{true}}\|_2^2/\|\boldsymbol{x}_{\text{true}}\|_2^2$, where $\boldsymbol{x}_{\text{true}}$ and $\widehat{\boldsymbol{x}}$ are the true and reconstructed signals, respectively. All iterative methods were initialized by FBP reconstructions implemented by IRT [49].

Figs. 2(b)–2(d) show reconstructions and corresponding RSE for one random realization of the noise and detector variation $\boldsymbol{c}$, with the expected total annihilation photon count (SNR) equal to 10^8; the optimal a is 0.5. All sparse reconstruction methods (PNPG, AT, CP, SPIRAL, and VMILA) perform similarly as long as they employ the same penalty: the TV sparsity penalty is superior to its DWT counterpart; see [9, Fig. 6] which shows average RSE of different methods as functions of $\mathbf{1}^T\Phi\boldsymbol{x}_{\text{true}}$.

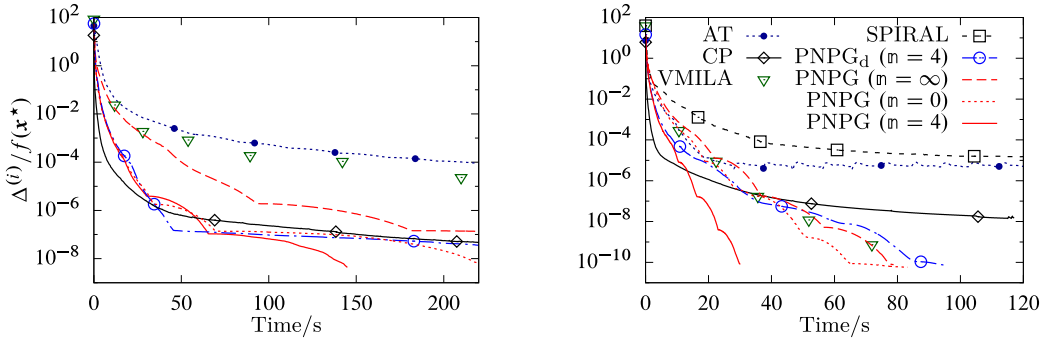
Figs. 3 and 4 show the normalized centered objectives $\Delta^{(i)}/f(\boldsymbol{x}^\star)$ as functions of the number of iterations and CPU time for the DWT and TV signal sparsity regularizations and two random realizations of the noise and detector variation with different total expected photon counts. The legends in Figs. 3(b) and 4(b) apply to Figs. 3(a) and 4(a) as well. Fig. 3 examines the convergence of PNPG as a function of the momentum tuning constants (γ, b) in (22), using $\gamma \in \{2, 5, 15\}$ and $b \in \{0, 1/4\}$. For small $\gamma \leq 5$, there is no significant difference between different selections and no choice is uniformly the best, consistent with [42] which considers only $b = 0$ and non-adaptive step size. As we increase γ further ($\gamma = 15$), we observe slower convergence. In the remainder of this section, we use $(\gamma, b) = (2, 1/4)$.

To illustrate the benefits of step-size adaptation, we present in Fig. 4 the performance of PNPG ($\mathtt{m} = \infty$), which does not adapt to the local curvature of the NLL and has monotonically non-increasing step sizes $\beta^{(i)}$, similar to FISTA. PNPG



(a) DWT, $\mathbf{1}^T\Phi\boldsymbol{x}_{\text{true}} = 10^8$, $a = 0.5$ (b) TV, $\mathbf{1}^T\Phi\boldsymbol{x}_{\text{true}} = 10^7$, $a = 0$

Fig. 3. Normalized centered objectives as functions of the number of iterations for (a) DWT and (b) TV regularizations.



(a) DWT, $\mathbf{1}^T\Phi\boldsymbol{x}_{\text{true}} = 10^8$, $a = 0.5$ (b) TV, $\mathbf{1}^T\Phi\boldsymbol{x}_{\text{true}} = 10^7$, $a = 0$

Fig. 4. Normalized centered objectives as functions of the central processing unit (CPU) time for (a) DWT and (b) TV regularizations.

($\mathbb{m} = 4$) outperforms PNPG ($\mathbb{m} = \infty$) because it uses step-size adaptation; see also Fig. 1(a), which corresponds to Fig. 4(a) and shows that the step size of PNPG ($\mathbb{m} = 4$) is consistently larger than that of PNPG ($\mathbb{m} = \infty$). Initializing PNPG iterations by a vector close to $\mathbf{0}$ (rather than with FBP) will lead to an even larger difference in convergence speed between PNPG ($\mathbb{m} = \infty$) and PNPG ($\mathbb{m} = 4$). The advantage of PNPG ($\mathbb{m} = 4$) over the aggressive PNPG ($\mathbb{m} = 0$) scheme is due to the *patient* nature of its step-size adaptation, which leads to a better local majorization function of the NLL and reduces time spent backtracking. Indeed, if we do not account for the time spent on each iteration and only compare the objectives as functions of the iteration index, then PNPG ($\mathbb{m} = 4$) and PNPG ($\mathbb{m} = 0$) perform similarly; see [8, Fig.4]. Although PNPG ($\mathbb{m} = 0$) and AT have the same step-size selection strategy and $\mathcal{O}(k^{-2})$ convergence-rate guarantees, PNPG ($\mathbb{m} = 0$) converges faster; both schemes are further outperformed by PNPG ($\mathbb{m} = 4$). Fig. 4(b) shows that SPIRAL, which does not employ PG step acceleration, takes at least three times longer than PNPG ($\mathbb{m} = 4$) to reach the same objective function.

In Fig. 4(b), AT and SPIRAL reach the performance floor due to their fixed inner convergence criterion in (41a); we observe a performance floor for AT in Fig. 4(a) as well. Reducing ϵ' in (41b) will lower this floor, at the cost of slowing down the two algorithms. This result justifies our convex-set projection in (20d) for the Nesterov's acceleration step, shows the superiority of (20d) over AT's acceleration in (39a) and (39c), and is consistent with the results in Section V-B.

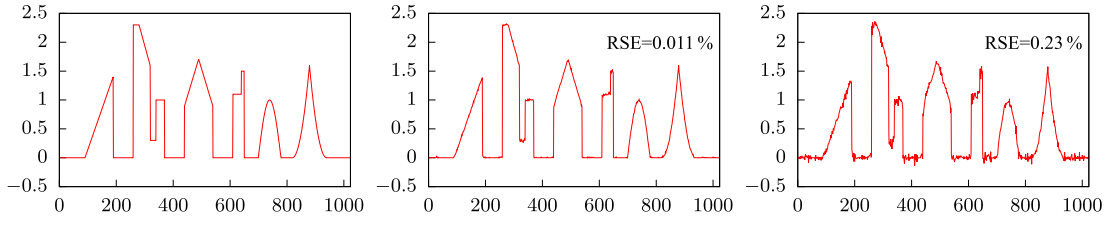
PNPG_{d} ($\mathbb{m} = 4$) employs the duality-gap–based inner convergence criterion (31b) with $q = 1.0001$. Since the goal of using (31b) is to guarantee (37), this inner criterion is more stringent and leads to slower overall performance of PNPG_{d} ($\mathbb{m} = 4$) compared to PNPG ($\mathbb{m} = 4$). Indeed, if we do not account for the time spent on each iteration and only compare the objectives as functions of the iteration index, then PNPG_{d} ($\mathbb{m} = 4$) and PNPG ($\mathbb{m} = 4$) perform similarly with the former slightly better.

The CP method uses the following tuning constants carefully selected for this particular problem: $(\sigma_1, \sigma_2, \tau) = (10^{-6}, 1, 10^{-3})$ and $(\sigma_1, \sigma_2, \tau) = (10^{-4}, 1, 10^{-2})$ for the DWT and TV penalties, respectively. CP is sensitive to tuning and a different selection of $(\sigma_1, \sigma_2, \tau)$ can significantly slow down its convergence. Initially, CP converges quickly and then slows down as it approaches the optimum.

Considering its $\mathcal{O}(k^{-1})$ theoretical convergence rate, VMILA performs quite well, thanks to its use of the variable-metric/scaling approach.

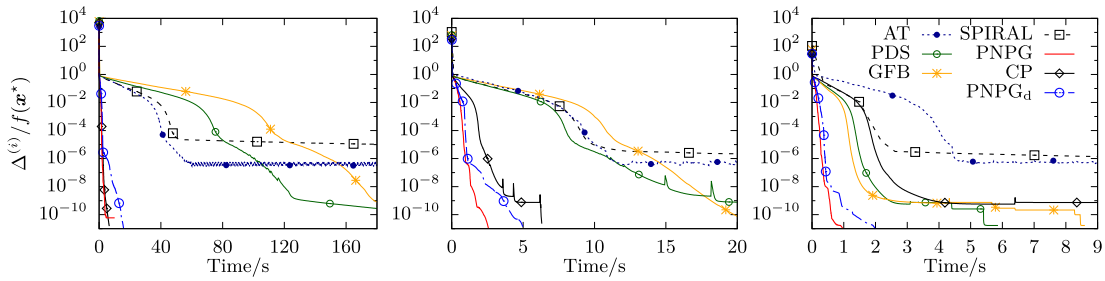
B. Skyline Signal Reconstruction From Linear Measurements

We adopt the DWT sparsifying transform and linear measurement model with Gaussian noise in Section II-B where the columns of the sensing matrix Φ are i.i.d. and drawn from the uniform distribution on unit sphere. Due to the widespread use of this measurement model, we can compare a wider range of methods than in the Poisson PET example in Section V-A.



(a) true signal (b) PNPG (c) NPG$_S$

Fig. 5. Nonnegative skyline signal and its PNPG and NPGS reconstructions for $N/p = 0.34$.



(a) $a = -5$ (b) $a = -4$ (c) $a = -3$

Fig. 6. Normalized centered objectives as functions of CPU time for normalized numbers of measurements $N/p = 0.34$ and different regularization constants a.

We have designed a "skyline" signal of length $p = 1024$ by overlapping magnified and shifted triangle, rectangle, sinusoid, and parabola functions; see Fig. 5(a). We generate the noiseless measurements using $\boldsymbol{y} = \Phi \boldsymbol{x}_{\text{true}}$. The DWT matrix Ψ is constructed using the Daubechies-4 wavelet with three decomposition levels whose approximation by the 5% largest-magnitude wavelet coefficients achieves RSE = 98%. We compare:

- AT, PNPG, and PNPG$_d$;
- CP with the parameters $\sigma_2 = \sigma_1$ [27] and $\tau = 1$ with σ_1 tuned separately for best performance in each experiment;[3]
- linearly constrained gradient projection method [5], part of the SPIRAL toolbox [46] and labeled SPIRAL herein;
- the GFB method [25] (see (4)):

$$z_1 \leftarrow z_1 + \lambda\big[\text{prox}_{(\gamma u/w)\rho}(2\boldsymbol{x} - z_1 - \gamma\nabla\mathcal{L}(\boldsymbol{x})) - \boldsymbol{x}\big] \tag{45a}$$

$$z_2 \leftarrow z_2 + \lambda[P_C(2\boldsymbol{x} - z_2 - \gamma\nabla\mathcal{L}(\boldsymbol{x})) - \boldsymbol{x}] \tag{45b}$$

$$\boldsymbol{x} \leftarrow wz_1 + (1-w)z_2 \tag{45c}$$

  with $\gamma = 1.9/\|\Phi\|_2^2$, $\lambda = 1$, and $w = 0.5$ tuned for best performance; and
- the PDS method [28]:

$$\bar{z} \leftarrow P_{[-u,u]^p}(z + \sigma\Psi^T\boldsymbol{x}) \tag{46a}$$

$$\bar{\boldsymbol{x}} \leftarrow P_C\left(\boldsymbol{x} - \tau\nabla\mathcal{L}(\boldsymbol{x}) - \tau\Psi(2\bar{z} - z)\right) \tag{46b}$$

$$z \leftarrow z + \gamma(\bar{z} - z) \tag{46c}$$

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \gamma(\bar{\boldsymbol{x}} - \boldsymbol{x}) \tag{46d}$$

[3]We select $\sigma_1 = \sigma_2$ as in [27] because $\|\Phi\|_2$ and $\|\Psi\|_2$ have approximately the same scale in this example.

where we choose $\tau = 1/(\sigma + \|\Phi\|_2^2/2)$ and $\gamma = 2 - 0.5\|\Phi\|_2^2(\tau^{-1} - \sigma)^{-1}$ with σ tuned for best performance,[4] all of which aim to solve the generalized analysis BPDN problem with a convex signal constraint. Here, $p' = p$, $\Psi \in \mathbb{R}^{p\times p}$ is an orthogonal matrix ($\Psi\Psi^T = \Psi^T\Psi = I$), and $\text{prox}_{\lambda\rho}\boldsymbol{a} = \Psi\mathcal{T}_\lambda(\Psi^T\boldsymbol{a})$ has a closed-form solution (see (6c)), which simplifies the implementation of the GFB method ((45a), in particular); see the discussion in Section I. The other tuning options for SPIRAL, and AT are kept to their default values, unless specified otherwise.

We initialize the iterative methods by the approximate minimum-norm estimate: $\boldsymbol{x}^{(0)} = \Phi^T[\text{E}(\Phi\Phi^T)]^{-1}\boldsymbol{y} = N\Phi^T\boldsymbol{y}/p$ and select the regularization parameter u as

$$u = 10^a U, \qquad U \triangleq \|\Psi^T\nabla\mathcal{L}(\boldsymbol{0})\|_\infty \tag{47}$$

where a is an integer selected from the interval $[-9, -1]$ and U is an upper bound on u of interest. Indeed, the minimum point $\boldsymbol{x}^\star$ reduces to $\boldsymbol{0}$ if $u \geq U$ [38, Sec. II-D].

As before, PNPG ($\mathtt{m} = 4$) and PNPG ($\mathtt{m} = 0$) converge at similar rates as functions of the number of iterations. However, due to the excessive attempts to increase the step size at every iteration, PNPG ($\mathtt{m} = 0$) spends more time backtracking and converges at a slower rate as a function of CPU time compared with PNPG ($\mathtt{m} = 4$); see also Fig. 1(b) which corresponds to Fig. 6(b) and shows the step sizes as functions of the number of iterations for $a = -4$ and $N/p = 0.34$. Hence, we present only the performances of PNPG with $\mathtt{m} = 4$ in this section.

Fig. 5 shows the advantage brought by the convex-set nonnegativity signal constraints (2). Figs. 5(b) and 5(c) present the

[4]These choices of τ and γ are at the boundary of the convergence region in [28, Th. 3.1]. We have searched for τ and γ inside this convergence region as well, but found that the boundary choices that we select are the best, or close to the best.

PNPG ($a = -5$) and NPGS ($a = -4$) reconstructions from one realization of the linear measurements with $N/p = 0.34$ and a tuned for the best RSE performance. Recall that NPGS imposes signal sparsity only. Here, imposing signal nonnegativity significantly improves the overall reconstruction and *does not* simply rectify the signal values close to zero.

Fig. 6 presents the normalized centered objectives $\Delta^{(i)}/f(\boldsymbol{x}^\star)$ as functions of CPU time for a random realization of the sensing matrix Φ with normalized numbers of measurements $N/p = 0.34$ and several different regularization constants a. (For the GFB method, we compute the normalized centered objectives using $P_C(\boldsymbol{x}^{(i)})$ instead of $\boldsymbol{x}^{(i)}$ in (34a) because its $\boldsymbol{x}^{(i)}$ may be outside C.) The legend in Fig. 6(c) applies also to Figs. 6(a) and 6(b). To achieve good performance, CP and PDS need to be manually tuned for each a. CP and PDS have optimal $\sigma_1 = \sigma_2$ equal to 0.01,0.1,1 and 0.0026, 0.026, 2.6 for $a = -5, -4, -3$, respectively.

PNPG and PNPG$_\mathrm{d}$ have the steepest descent rate, followed by PNPG$_\mathrm{d}$. AT and SPIRAL reach the performance floor around the relative precision of 10^{-6} due to their fixed inner convergence criterion in (41a). The GFB and primal-dual methods, PDS and CP, are sensitive to the selection of the tuning constants. After a careful selection of the tuning constants, CP performs exceptionally well in Figs. 6(a) and 6(b). The performance of GFB is affected significantly by the value of the regularization parameter a.

VI. Conclusion

We developed a fast algorithm for reconstructing sparse signals that belong to a closed convex set by employing a projected proximal-gradient scheme with Nesterov's acceleration, restart, and adaptive step size. We applied the PNPG method to construct one of the first Nesterov-accelerated proximal-gradient reconstruction algorithm for Poisson compressed sensing. We presented integrated derivation of the proposed algorithm and convergence-rate upper bound that accounts for inexactness of the proximal operator and also proved convergence of iterates. Our PNPG approach is computationally efficient compared with other state-of-the-art methods.

Appendix A
Derivation of Acceleration (20a)–(20d) and Proofs of Lemma 1 and Theorem 1

We first prove Lemma 1 and then derive the acceleration (20a)–(20d) and prove Theorem 1.

Proof of Lemma 1: According to Definition 1 and (24),

$$ur(\boldsymbol{x}) \geq ur(\boldsymbol{x}^{(i)}) + (\boldsymbol{x} - \boldsymbol{x}^{(i)})^T \left[\frac{\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i)}}{\beta^{(i)}} - \nabla \mathcal{L}(\bar{\boldsymbol{x}}^{(i)}) \right] - \frac{(\varepsilon^{(i)})^2}{2\beta^{(i)}} \tag{A1a}$$

for any $\boldsymbol{x} \in \mathbb{R}^p$. Moreover, due to the convexity of $\mathcal{L}(\boldsymbol{x})$,

$$\mathcal{L}(\boldsymbol{x}) \geq \mathcal{L}(\bar{\boldsymbol{x}}^{(i)}) + (\boldsymbol{x} - \bar{\boldsymbol{x}}^{(i)})^T \nabla \mathcal{L}(\bar{\boldsymbol{x}}^{(i)}). \tag{A1b}$$

Summing (A1a), (A1b), and (21) completes the proof. ■

The following result from [34, Prop. 3.2.1 in Sec. 3.2] states that the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ can be reduced by projecting them onto a closed convex set C.

Lemma 2 (Projection theorem): The projection mapping onto a nonempty closed convex set $C \subseteq \mathbb{R}^p$ is nonexpansive

$$\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|_2^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{A2}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$.

We now derive the projected Nesterov's acceleration step (20b)–(20d) with the goal of selecting the $\bar{\boldsymbol{x}}^{(i)}$ in the proximal step (20e) that achieves the convergence rate of $\mathcal{O}(k^{-2})$. This derivation and convergence-rate proof are inspired by—but are more general than—[16]. We start from (25) with $\boldsymbol{x}$ replaced by $\boldsymbol{x} = \boldsymbol{x}^\star$ and $\boldsymbol{x} = \boldsymbol{x}^{(i-1)}$,

$$-\Delta^{(i)} \geq \frac{\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^\star\|_2^2 - \|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^\star\|_2^2 - (\varepsilon^{(i)})^2}{2\beta^{(i)}} \tag{A3a}$$

$$\Delta^{(i-1)} - \Delta^{(i)} \geq \frac{\delta^{(i)} - \|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^{(i-1)}\|_2^2 - (\varepsilon^{(i)})^2}{2\beta^{(i)}} \tag{A3b}$$

and design two coefficient sequences $a^{(i)} > 0$ and $b^{(i)} > 0$ that multiply (A3a) and (A3b), respectively, which ultimately leads to (20a)–(20d) and the convergence-rate guarantee in (33a).

Consider sequences $a^{(i)} > 0$ and $b^{(i)} > 0$. Multiply them by (A3a) and (A3b), respectively, add the resulting expressions, and multiply by $\beta^{(i)}$ to obtain

$$\begin{aligned} &-2\beta^{(i)}c^{(i)}\Delta^{(i)} + 2\beta^{(i)}b^{(i)}\Delta^{(i-1)} \\ &\geq \frac{1}{c^{(i)}}\left\|c^{(i)}\boldsymbol{x}^{(i)} - b^{(i)}\boldsymbol{x}^{(i-1)} - a^{(i)}\boldsymbol{x}^\star\right\|_2^2 \\ &\quad - \frac{1}{c^{(i)}}\|c^{(i)}\bar{\boldsymbol{x}}^{(i)} - b^{(i)}\boldsymbol{x}^{(i-1)} - a^{(i)}\boldsymbol{x}^\star\|_2^2 - c^{(i)}(\varepsilon^{(i)})^2 \\ &= c^{(i)}\left[t^{(i)} - \bar{t}^{(i)} - (\varepsilon^{(i)})^2\right] \end{aligned} \tag{A4}$$

where

$$c^{(i)} \triangleq a^{(i)} + b^{(i)} \tag{A5a}$$

$$t^{(i)} \triangleq \|\boldsymbol{x}^{(i)} - \boldsymbol{z}^{(i)}\|_2^2, \quad \bar{t}^{(i)} \triangleq \|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{z}^{(i)}\|_2^2 \tag{A5b}$$

$$\boldsymbol{z}^{(i)} \triangleq \frac{b^{(i)}}{c^{(i)}}\boldsymbol{x}^{(i-1)} + \frac{a^{(i)}}{c^{(i)}}\boldsymbol{x}^\star. \tag{A5c}$$

We arranged (A4) using completion of squares so that the first two summands are similar (but with opposite signs), with the goal of facilitating cancellations as we sum over i. Since we have control over the sequences $a^{(i)}$ and $b^{(i)}$, we impose the following boundary conditions for $i \geq 1$:

$$c^{(i-1)}t^{(i-1)} \geq c^{(i)}\bar{t}^{(i)} \tag{A6a}$$

$$\pi^{(i)} \geq 0 \tag{A6a}$$

where

$$\pi^{(i)} \triangleq \beta^{(i)}c^{(i)} - \beta^{(i+1)}b^{(i+1)}. \tag{A7}$$

Now, apply the inequality (A6a) to the right-hand side of (A4):

$$-2\beta^{(i)}c^{(i)}\Delta^{(i)} + 2\beta^{(i)}b^{(i)}\Delta^{(i-1)} \geq c^{(i)}t^{(i)} - c^{(i-1)}t^{(i-1)} - c^{(i)}(\varepsilon^{(i)})^2 \tag{A8a}$$

and sum (A8a) over $i = 1, 2, \ldots, k$, which leads to summand cancellations and

$$-2\beta^{(k)}c^{(k)}\Delta^{(k)} + 2\beta^{(1)}b^{(1)}\Delta^{(0)} - 2\sum_{i=1}^{k-1}\pi^{(i)}\Delta^{(i)}$$

$$\geq -c^{(0)}t^{(0)} - \sum_{i=1}^{k} c^{(i)}(\varepsilon^{(i)})^2 \tag{A8b}$$

where (A8b) follows by discarding a nonnegative term $c^{(k)}t^{(k)}$.

Now, due to $\pi^{(i)}\Delta^{(i)} \geq 0$ (see (34a) and (A6b)), the inequality (A8b) leads to

$$\Delta^{(k)} \leq \frac{2\beta^{(1)}b^{(1)}\Delta^{(0)} + c^{(0)}t^{(0)} + \sum_{i=1}^{k} c^{(i)}(\varepsilon^{(i)})^2}{2\beta^{(k)}c^{(k)}} \tag{A9}$$

with simple upper bound on the right-hand side, thanks to summand cancellations facilitated by the assumptions (A6).

As long as $\beta^{(k)}c^{(k)}$ grows at a rate of k^2 and the inexactness of the proximal mappings leads to bounded $\sum_{i=1}^{k} c^{(i)}(\varepsilon^{(i)})^2$, the centered objective function $\Delta^{(k)}$ can achieve the desired bound decrease rate of $1/k^2$.

In the following section, we show how to satisfy (A6a), which will lead to the projected momentum acceleration step (20d). We approach the constraints (A6a) by first aiming to meet them with equality, which is possible in the absence of the convex-set constraint ($C = \mathbb{R}^p$). We then use the nonexpansiveness of the convex-set projection to construct $a^{(i)}$ and $b^{(i)}$ that satisfy (A6a) with inequality in the general case where the convex-set constraint is present. Finally, we show how to satisfy (A6b), which will allow us to construct the recursive update of $\theta^{(i)}$ in (20b) and verify the allowed range of momentum tuning constants in (22).

I. Satisfying Conditions (A6)

a) Imposing Equality in (A6a): (A6a) holds with equality for all i and any $\boldsymbol{x}^\star$ when we choose $\bar{\boldsymbol{x}}^{(i)} = \widehat{\boldsymbol{x}}^{(i)}$ that satisfies

$$\sqrt{c^{(i-1)}}(\boldsymbol{x}^{(i-1)} - \boldsymbol{z}^{(i-1)}) = \sqrt{c^{(i)}}(\widehat{\boldsymbol{x}}^{(i)} - \boldsymbol{z}^{(i)}). \tag{A10}$$

Now, (A10) requires equal coefficients multiplying $\boldsymbol{x}^\star$ on both sides; thus $a^{(i)}/\sqrt{c^{(i)}} = 1/\mathrm{w}$ for all i, where $\mathrm{w} > 0$ is a constant (not a function of i), which implies $c^{(i)} = \mathrm{w}^2(a^{(i)})^2$ and $b^{(i)} = \mathrm{w}^2(a^{(i)})^2 - a^{(i)}$; see also (A5a). Upon defining

$$\theta^{(i)} \triangleq \mathrm{w}^2 a^{(i)} \tag{A11a}$$

we have

$$\mathrm{w}^2 c^{(i)} = (\theta^{(i)})^2; \quad \mathrm{w}^2 b^{(i)} = (\theta^{(i)})^2 - \theta^{(i)}. \tag{A11b}$$

Plug (A11) into (A10) and reorganize to obtain the following form of momentum acceleration:

$$\widehat{\boldsymbol{x}}^{(i)} = \boldsymbol{x}^{(i-1)} + \Theta^{(i)}\left(\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^{(i-2)}\right). \tag{A12}$$

Although $\bar{\boldsymbol{x}}^{(i)} = \widehat{\boldsymbol{x}}^{(i)}$ satisfies (A6a), it is not guaranteed to be within $\mathrm{dom}\,\mathcal{L}$; consequently, the proximal-mapping step for this selection *may not* be computable.

b) Selecting $\bar{\boldsymbol{x}}^{(i)} \in C$ That Satisfies (A6a): We now seek $\bar{\boldsymbol{x}}^{(i)}$ within C that satisfies the inequality (A6a). Since $\boldsymbol{x}^{(i-1)}$ and $\boldsymbol{x}^\star$ are in C, $\boldsymbol{z}^{(i)} \in C$ by the convexity of C; see (A5c). According to Lemma 2, projecting (A12) onto C preserves or reduces the distance between points. Therefore,

$$\bar{\boldsymbol{x}}^{(i)} = P_C(\widehat{\boldsymbol{x}}^{(i)}) \tag{A13}$$

belongs to C *and* satisfies the condition (A6a):

$$c^{(i-1)}t^{(i-1)} = c^{(i)}\|\widehat{\boldsymbol{x}}^{(i)} - \boldsymbol{z}^{(i)}\|_2^2 \tag{A14a}$$

$$\geq c^{(i)}\|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{z}^{(i)}\|_2^2 = c^{(i)}\bar{t}^{(i)} \tag{A14b}$$

where (A14a) and (A14b) follow from (A10) and by using Lemma 2, respectively; see also (A5b).

Without loss of generality, set $\mathrm{w} = 1$ and rewrite and modify (A7), (A5b), and (A8b) using (A11) to obtain

$$\pi^{(i)} = \beta^{(i)}(\theta^{(i)})^2 - \beta^{(i+1)}\theta^{(i+1)}(\theta^{(i+1)} - 1), \quad i \geq 1 \tag{A15a}$$

$$(\theta^{(i)})^2 t^{(i)} = \left\|\theta^{(i)}\boldsymbol{x}^{(i)} - (\theta^{(i)} - 1)\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^\star\right\|_2^2 \tag{A15b}$$

$$\sum_{i=1}^{k-1}\pi^{(i)}\Delta^{(i)} \leq \frac{1}{2}\left[(\theta^{(0)})^2 t^{(0)} + \sum_{i=1}^{k}(\theta^{(i)}\varepsilon^{(i)})^2\right] \tag{A15c}$$

where (A15c) is obtained by discarding the negative term $-2\beta^{(k)}(\theta^{(k)})^2\Delta^{(k)}$ and the zero term $\beta^{(1)}\theta^{(1)}(\theta^{(1)} - 1)\Delta^{(0)}$ (because $\theta^{(1)} = 1$) on the left-hand side of (A8b). Now, (33a) follows from (A9) by using $\theta^{(0)} = \theta^{(1)} = 1$ (see (20b)), (A11), and (A15b) with $i = 0$.

c) Satisfying (A6b): By substituting (A15a) into (A6b), we obtain the conditions

$$\beta^{(i-1)}(\theta^{(i-1)})^2 \geq \beta^{(i)}\left[(\theta^{(i)})^2 - \theta^{(i)}\right] \tag{A16a}$$

and interpret $(\pi^{(i)})_{i=1}^{+\infty}$ as the sequence of gaps between the two sides of (A16a); (A16a) implies

$$\theta^{(i)} \leq 1/2 + \sqrt{1/4 + B^{(i)}(\theta^{(i-1)})^2}. \tag{A16b}$$

Comparing (20b) with (A16b) justifies the constraints in (22).

II. Connection to Convergence-Rate Analysis of FISTA in [16]

If the step-size sequence $(\beta^{(i)})$ is non-increasing (e.g., in the backtracking-only scenario with $\mathbb{n} = +\infty$), (20b) with $B^{(i)} = 1$ also satisfies the inequality (A16b). In this case, (33a) still holds but (33b) does not because (35) no longer holds. However, because $B^{(i)} = 1$, we have $\theta^{(k)} \geq (k+1)/\gamma$ and

$$\Delta^{(k)} \leq \gamma^2 \frac{\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2 + \mathcal{E}^{(k)}}{2\beta^{(k)}(k+1)^2} \tag{A17}$$

which generalizes [16, Th. 4.4] to include the inexactness of the proximal operator and the convex-set projection.

Appendix B
Convergence of Iterates

To prove convergence of iterates, we need to show that the centered objective function $\Delta^{(k)}$ decreases faster than the right-hand side of (33b). We introduce Lemmas 3 and 4 and

then use them to prove Theorem 2. Throughout this Appendix, we assume that Assumption 1 of Theorem 2 holds, which justifies (A3) and (A16) as well as results from Appendix A that we use in the proofs.

Lemma 3: Under Assumptions 1–3 of Theorem 2,

$$\sum_{i=1}^{+\infty}(2\theta^{(i)}-1)\delta^{(i)} < +\infty. \tag{B1}$$

Proof: By letting $k \to +\infty$ in (A15c) and using (37), we obtain

$$\sum_{i=1}^{+\infty}\pi^{(i)}\Delta^{(i)} < +\infty. \tag{B2}$$

For $i \geq 1$, rewrite (A15a) using $\theta^{(i)}$ expressed in terms of $\theta^{(i+1)}$ (based on (20b)):

$$\begin{aligned}\pi^{(i)} &= \frac{\beta^{(i+1)}}{\gamma}\big[(\gamma-2)\theta^{(i+1)} + (1-b\gamma^2)/\gamma\big] \\ &\geq \frac{\gamma-2}{\gamma}\beta^{(i+1)}\theta^{(i+1)} \end{aligned}\tag{B3}$$

where the inequality in (B3) is due to $b\gamma^2 - 1 < 0$; see Assumption 3. Apply nonexpansiveness of the projection operator to (A3b) and use (A12) to obtain

$$2\beta^{(i)}(\Delta^{(i-1)} - \Delta^{(i)}) \geq \delta^{(i)} - (\Theta^{(i)})^2\delta^{(i-1)} - (\varepsilon^{(i)})^2; \tag{B4}$$

then multiply both sides of (B4) by $(\theta^{(i)})^2$, sum over $i = 1, 2, \ldots, k$ and reorganize:

$$\begin{aligned}\sum_{i=1}^{k-1}(2\theta^{(i)}-1)\delta^{(i)} &\leq (\theta^{(0)}-1)^2\delta^{(0)} - (\theta^{(k)})^2\delta^{(k)} + \mathcal{E}^{(k)} \\ &+ 2\beta^{(1)}\Delta^{(0)} - 2\beta^{(k)}(\theta^{(k)})^2\Delta^{(k)} + 2\sum_{i=1}^{k-1}\varrho^{(i)}\Delta^{(i)} \qquad \text{(B5a)} \\ &\leq \mathcal{E}^{(k)} + 2\beta^{(1)}\Delta^{(0)} + \frac{4}{\gamma-2}\sum_{i=1}^{k-1}\pi^{(i)}\Delta^{(i)} \qquad \text{(B5b)}\end{aligned}$$

where (see (A15a))

$$\begin{aligned}\varrho^{(i)} &\triangleq \beta^{(i+1)}(\theta^{(i+1)})^2 - \beta^{(i)}(\theta^{(i)})^2 \qquad \text{(B5c)} \\ &= \beta^{(i+1)}\theta^{(i+1)} - \pi^{(i)}, \qquad \text{(B5d)}\end{aligned}$$

and we drop the zero term $(\theta^{(0)}-1)^2\delta^{(0)}$ and the negative term $-(\theta^{(k)})^2\delta^{(k)} - 2\beta^{(k)}(\theta^{(k)})^2\Delta^{(k)}$ from (B5a) and use the fact that $\varrho^{(i)} \leq [2/(\gamma-2)]\pi^{(i)}$ implied by (B3) to obtain (B5b). Finally, let $k \to +\infty$ and use (37) and (B2) to conclude (B1). ■

Lemma 4: For $j \geq 3$,

$$\Pi_j \triangleq \sum_{k=j}^{+\infty}\prod_{\ell=j}^{k}\Theta^{(\ell)} \leq \gamma\theta^{(j-1)} - 1. \tag{B6}$$

Proof: For $j \geq 3$,

$$\begin{aligned}\frac{1}{\sqrt{\beta^{(k-1)}}\theta^{(k-1)}\theta^{(k)}} &\leq \frac{\gamma}{\sqrt{\beta^{(k-1)}}\theta^{(k-1)}} - \frac{\gamma}{\sqrt{\beta^{(k)}}\theta^{(k)}} \qquad \text{(B7a)} \\ &\leq \frac{\gamma}{\sqrt{\beta^{(k-2)}}\theta^{(k-2)}} - \frac{\gamma}{\sqrt{\beta^{(k)}}\theta^{(k)}} \qquad \text{(B7b)}\end{aligned}$$

where we obtain the inequality (B7a) by combining the terms on the right-hand size and using (35a) and (B7b) holds because $\sqrt{\beta^{(k)}}\theta^{(k)}$ is an increasing sequence (see Section IV). Now,

$$\begin{aligned}\Pi_j &\leq \sum_{k=j}^{+\infty}\prod_{\ell=j}^{k}\frac{\beta^{(\ell-2)}(\theta^{(\ell-2)})^2}{\beta^{(\ell-1)}\theta^{(\ell-1)}\theta^{(\ell)}} = \sum_{k=j}^{+\infty}\frac{\beta^{(j-2)}(\theta^{(j-2)})^2\theta^{(j-1)}}{\beta^{(k-1)}(\theta^{(k-1)})^2\theta^{(k)}} \qquad \text{(B8a)} \\ &\leq \frac{\gamma\beta^{(j-2)}(\theta^{(j-2)})^2\theta^{(j-1)}}{\sqrt{\beta^{(j-2)}}\theta^{(j-2)}\sqrt{\beta^{(j-1)}}\theta^{(j-1)}} = \gamma\sqrt{B^{(j-1)}}\theta^{(j-2)} \qquad \text{(B8b)}\end{aligned}$$

where (B8a) follows by using (20c), (A16a) with $i = \ell - 1$, and fraction-term cancellation; (B8b) is obtained by substituting (B7b) into (B8a) and canceling summation terms. (B8b) implies (B6) by using (35a) with $k = j - 1$. ■

Define

$$\lambda^{(i)} \triangleq \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^\star\|_2^2, \quad \Lambda^{(i)} \triangleq \lambda^{(i)} - \lambda^{(i-1)}. \tag{B9}$$

Since $f(\boldsymbol{x}^{(i)})$ converges to $f(\boldsymbol{x}^\star) = \min_{\boldsymbol{x}} f(\boldsymbol{x})$ as the iteration index i grows and $\boldsymbol{x}^\star$ is a minimizer, it is sufficient to prove the convergence of $\lambda^{(i)}$; see [42, Th. 4.1].

Proof of Theorem 2: Use (A3a) and $\Delta^{(i)} \geq 0$ to obtain

$$0 \geq \lambda^{(i)} - \|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^\star\|_2^2 - (\varepsilon^{(i)})^2. \tag{B10}$$

Now,

$$\begin{aligned}\|\bar{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^\star\|_2^2 &\leq \|\widehat{\boldsymbol{x}}^{(i)} - \boldsymbol{x}^\star\|_2^2 = \lambda^{(i-1)} + (\Theta^{(i)})^2\delta^{(i-1)} \\ &+ 2\Theta^{(i)}(\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^\star)^T(\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^{(i-2)}) \qquad \text{(B11a)} \\ &\leq \lambda^{(i-1)} + (\Theta^{(i)})^2\delta^{(i-1)} + \Theta^{(i)}(\Lambda^{(i-1)} + \delta^{(i-1)}) \qquad \text{(B11b)}\end{aligned}$$

where (B11a) and (B11b) follow by using the nonexpansiveness of the projection operator (see also (A12)) and the identity

$$2(\boldsymbol{a}-\boldsymbol{b})^T(\boldsymbol{a}-\boldsymbol{c}) = \|\boldsymbol{a}-\boldsymbol{b}\|_2^2 + \|\boldsymbol{a}-\boldsymbol{c}\|_2^2 - \|\boldsymbol{b}-\boldsymbol{c}\|_2^2 \tag{B12}$$

respectively. Combine the inequalities (B11b) and (B10) to get

$$\begin{aligned}\Lambda^{(i)} &\leq \Theta^{(i)}\big[\Lambda^{(i-1)} + (\Theta^{(i)}+1)\delta^{(i-1)}\big] + (\varepsilon^{(i)})^2 \qquad \text{(B13a)} \\ &\leq \Theta^{(i)}(\Lambda^{(i-1)} + 2\delta^{(i-1)}/\xi) + (\varepsilon^{(i)})^2 \qquad \text{(B13b)}\end{aligned}$$

where (B13b) is due to $1 < 1/\xi$ (see (29)) and the following:

$$\begin{aligned}\Theta^{(i)} &< \frac{\theta^{(i-1)}}{\theta^{(i)}} = \frac{\sqrt{\beta^{(i-1)}}\theta^{(i-1)}\sqrt{\beta^{(i)}}}{\sqrt{\beta^{(i)}}\theta^{(i)}\sqrt{\beta^{(i-1)}}} \qquad \text{(B14a)} \\ &< \frac{\sqrt{\beta^{(i)}}}{\sqrt{\beta^{(i-1)}}} \leq \frac{1}{\sqrt{\xi}} < \frac{1}{\xi} \qquad \text{(B14b)}\end{aligned}$$

where we have used (20c), the fact that $\sqrt{\beta^{(i)}}\theta^{(i)}$ is an increasing sequence, $\beta^{(i)}/\beta^{(i-1)} \geq 1/\xi$ (see Section III-B), and (29).

According to (35b) and the fact that the sequence $(\beta^{(i)})$ is bounded (by Assumption 4), there exists an integer J such that

$$\theta^{(j-1)} \geq 2, \qquad \Theta^{(j)} \geq \frac{1}{\theta^{(j)}} > 0 \tag{B15}$$

for all $j \geq J$, where the second inequality follows from the first and the definition of $\Theta^{(j)}$; see (20c). Then

$$\begin{aligned}\Omega^{(i)} &\triangleq \max(0, \Lambda^{(i)}) \\ &\leq \Theta^{(i)} \left[\Omega^{(i-1)} + \frac{2\delta^{(i-1)}}{\xi} + \frac{(\varepsilon^{(i)})^2}{\Theta^{(i)}} \right] \qquad \text{(B16a)} \\ &\leq \sum_{j=J}^{i} \left[\frac{2\delta^{(j-1)}}{\xi} + \frac{(\varepsilon^{(j)})^2}{\Theta^{(j)}} \right] \prod_{\ell=j}^{i} \Theta^{(\ell)} \\ &\quad + \Omega^{(J-1)} \prod_{\ell=J}^{i} \Theta^{(\ell)} \qquad \text{(B16b)}\end{aligned}$$

for $i \geq J$, where the inequality in (B16a) follows by combining the inequalities (B13b) and $\Omega^{(i-1)} \geq \Lambda^{(i-1)}$, and (B16b) follows by recursively applying inequality (B16a) with i replaced by $i-1, i-2, \ldots, J$. Now, sum the inequalities (B16b) over $i = J, J+1, \ldots, +\infty$ and exchange the order of summation over i and j on the right-hand side (see also (B6)):

$$\sum_{i=J}^{+\infty} \Omega^{(i)} \leq \sum_{j=J}^{+\infty} \Pi_j \left[\frac{2\delta^{(j-1)}}{\xi} + \frac{(\varepsilon^{(j)})^2}{\Theta^{(j)}} \right] + \Pi_J \Omega^{(J-1)}. \tag{B17}$$

For $j \geq J \geq 3$,

$$\gamma(2\theta^{(j-1)} - 1) - \Pi_j \geq \gamma(\theta^{(j-1)} - 1) + 1 > 0 \tag{B18a}$$

$$2\gamma(\theta^{(j-1)} - 1) - \Pi_j \geq \gamma(\theta^{(j-1)} - 2) + 1 > 0 \tag{B18b}$$

where the first and second inequalities in (B18) follow by applying Lemma 4 and (B15), respectively; consequently,

$$\begin{aligned}\sum_{j=J}^{+\infty} \Pi_j \delta^{(j-1)} &\leq \gamma \sum_{j=J}^{+\infty} (2\theta^{(j)} - 1)\delta^{(j)} < +\infty \qquad \text{(B19a)} \\ \sum_{j=J}^{+\infty} \Pi_j \frac{(\varepsilon^{(j)})^2}{\Theta^{(j)}} &\leq 2\gamma \sum_{j=J}^{+\infty} (\varepsilon^{(j)})^2 \frac{\theta^{(j-1)} - 1}{\Theta^{(j)}} \qquad \text{(B19b)} \\ &= 2\gamma \sum_{j=J}^{+\infty} (\varepsilon^{(j)})^2 \theta^{(j)} \qquad \text{(B19c)} \\ &\leq 2\gamma \sum_{j=J}^{+\infty} (\theta^{(j)} \varepsilon^{(j)})^2 \qquad \text{(B19d)}\end{aligned}$$

where (B19a) follows from (B18a) and Lemma 3 (for the second inequality) and (B19b) follows by using (B18b); (B19c) and (B19d) are due to (20c) and (B15), respectively. Combine (B19a) and (B19d) with (B17) to conclude that

$$\sum_{i=1}^{+\infty} \Omega^{(i)} < +\infty. \tag{B20}$$

The remainder of the proof uses the technique employed by Chambolle and Dossal to conclude the proof of [42, Th. 4.1, p. 978], which we repeat for completeness. Define $X^{(i)} \triangleq \lambda^{(i)} - \sum_{j=1}^{i} \Omega^{(j)}$, which is lower bounded because $\lambda^{(i)}$ and $\sum_{j=1}^{i} \Omega^{(j)}$ are lower and upper bounded, respectively; see (B9) and (B20). Furthermore, $(X^{(i)})$ is an non-increasing sequence:

$$X^{(i+1)} = \lambda^{(i+1)} - \Omega^{(i+1)} - \sum_{j=1}^{i} \Omega^{(j)} \leq X^{(i)}, \tag{B21}$$

where we used the fact that $\Omega^{(i+1)} \geq \Lambda^{(i+1)} = \lambda^{(i+1)} - \lambda^{(i)}$. Hence, $(X^{(i)})$ converges as $i \to +\infty$. Since $\sum_{j=1}^{i} \Omega^{(j)}$ converges, $(\lambda^{(i)})$ also converges. ■

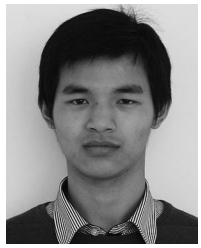
ACKNOWLEDGMENT

The authors would like to thank Prof. M. Hong, Iowa State University, for valuable discussions and the anonymous reviewers for their helpful comments.

REFERENCES

[1] S. Boyd and L. Vandenberghe, "Vectors, matrices, and least squares," 2016. [Online]. Available: http://stanford.edu/class/ee103/mma.pdf
[2] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
[3] J. L. Prince and J. M. Links, *Medical Imaging Signals and Systems*, 2nd ed. Upper Saddle River, NJ, USA: Pearson, 2015.
[4] Z. T. Harmany, R. F. Marcia, and R. M. Willett, "This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms—Theory and practice," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1084–1096, Mar. 2012.
[5] Z. Harmany, D. Thompson, R. Willett, and R. F. Marcia, "Gradient projection for linearly constrained convex optimization in sparse signal recovery," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, China, Sep. 2010, pp. 3361–3364.
[6] D. C. Youla and H. Webb, "Image restoration by the method of convex projections: Part 1—Theory," *IEEE Trans. Med. Imag.*, vol. MI-1, no. 2, pp. 81–94, Oct. 1982.
[7] B. E. Hansen and J. S. Racine, "Jackknife model averaging," *J. Econometrics*, vol. 167, no. 1, pp. 38–46, 2012.
[8] R. Gu and A. Dogandžić, "Projected Nesterov's proximal-gradient signal recovery from compressive Poisson measurements," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2015, pp. 1490–1495.
[9] R. Gu and A. Dogandžić, "Projected Nesterov's proximal-gradient algorithm for sparse signal reconstruction with a convex constraint," Oct. 2016, arXiv: 1502.02613.
[10] S. Bonettini, F. Porta, and V. Ruggiero, "A variable metric forward-backward method with extrapolation," *SIAM J. Sci. Comput.*, vol. 38, no. 4, pp. A2558–A2584, 2016.
[11] F.-X. Dupé, J. M. Fadili, and J.-L. Starck, "Deconvolution under Poisson noise using exact data fidelity and synthesis or analysis sparsity priors," *Statist. Methodol.*, vol. 9, no. 1–2, pp. 4–18, 2012.
[12] F.-X. Dupé, J. M. Fadili, and J.-L. Starck, "A proximal iteration for deconvolving Poisson noisy images using sparse representations," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 310–321, Feb. 2009.
[13] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Adv. Neural Inf. Process. Syst. 24*, 2011, pp. 1458–1466.
[14] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, "Accelerated and inexact forward-backward algorithms," *SIAM J. Optim.*, vol. 23, no. 3, pp. 1607–1633, 2013.
[15] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Math. Dokl.*, vol. 27, no. 2, pp. 372–376, 1983.
[16] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[17] Y. Nesterov, "Gradient methods for minimizing composite functions," *Math. Program., Ser. B*, vol. 140, no. 1, pp. 125–161, 2013.

[18] S. R. Becker, E. J. Candès, and M. C. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Math. Program. Comput.*, vol. 3, no. 3, pp. 165–218, 2011. [Online]. Available: http://cvxr.com/tfocs
[19] S. Bonettini, I. Loris, F. Porta, and M. Prato, "Variable metric inexact line-search-based methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 26, no. 2, pp. 891–921, 2016.
[20] A. Auslender and M. Teboulle, "Interior gradient and proximal methods for convex and conic optimization," *SIAM J. Optim.*, vol. 16, no. 3, pp. 697–725, 2006.
[21] J. Y. Bello Cruz and T. T. A. Nghia, "On the convergence of the forward—backward splitting method with linesearches," *Optim. Method. Softw.*, vol. 31, no. 6, pp. 1209–1238, 2016.
[22] C. Chaux, J.-C. Pesquet, and N. Pustelnik, "Nested iterative algorithms for convex constrained image recovery problems," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 730–762, 2009.
[23] S. Anthoine, J.-F. Aujol, Y. Boursier, and C. Mélot, "Some proximal methods for Poisson intensity CBCT and PET," *Inverse Probl. Imag.*, vol. 6, no. 4, pp. 565–598, 2012.
[24] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, Nov. 2009.
[25] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1199–1226, 2013.
[26] N. Pustelnik, C. Chaux, and J.-C. Pesquet, "Parallel proximal algorithm for image restoration using hybrid regularization," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2450–2462, Sep. 2011.
[27] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, 2011.
[28] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.
[29] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Adv. Comput. Math.*, vol. 38, no. 3, pp. 667–681, 2013.
[30] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds. New York, NY, USA: Springer, vol. 49. 2011, pp. 185–212.
[31] J. Liang, J. Fadili, and G. Peyré, "Convergence rates with inexact nonexpansive operators," *Math. Program., Ser. A*, vol. 159, no. 1, pp. 403–434, 2016.
[32] D. Davis, "Convergence rate analysis of primal-dual splitting schemes," *SIAM J. Optim.*, vol. 25, no. 3, pp. 1912–1943, 2015.
[33] S. Salzo, "The variable metric forward-backward splitting algorithm under mild differentiability assumptions," May 2016, arXiv: 1605.00952.
[34] D. P. Bertsekas, *Convex Optimization Algorithms*. Belmont, MA, USA: Athena Scientific, 2015.
[35] J. M. Ollinger and J. A. Fessler, "Positron-emission tomography," *IEEE Signal Process. Mag.*, vol. SPM-14, no. 1, pp. 43–55, Jan. 1997.
[36] L. Zanni, A. Benfenati, M. Bertero, and V. Ruggiero, "Numerical methods for parameter estimation in Poisson data inversion," *J. Math. Imag. Vis.*, vol. 52, no. 3, pp. 397–413, 2015.
[37] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed. New York, NY, USA: Chapman & Hall, 1989.
[38] R. Gu and A. Dogandžić, "Reconstruction of nonnegative sparse signals using accelerated proximal-gradient algorithms," Mar. 2015, arXiv: 1502.02613v3.
[39] R. Gu and A. Dogandžić, "Blind X-ray CT image reconstruction from polychromatic Poisson measurements," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 150–165, Jun. 2016.
[40] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes," *Found. Comput. Math.*, vol. 15, no. 3, pp. 715–732, 2015.
[41] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.
[42] A. Chambolle and C. Dossal, "On the convergence of the iterates of the 'fast iterative shrinkage/thresholding algorithm'," *J. Optim. Theory Appl.*, vol. 166, no. 3, pp. 968–982, 2015.
[43] J.-F. Aujol and C. Dossal, "Stability of over-relaxations for the forward-backward algorithm, application to FISTA," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2408–2433, 2015.
[44] I. Daubechies, M. Defrise, and C. De Mol, "Sparsity-enforcing regularisation and ISTA revisited," *Inverse Probl.*, vol. 32, no. 10, EID 104001, 2016.
[45] P. Tseng, "A modified forward-backward splitting method for maximal monotone mappings," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 431–446, 2000.
[46] Z. Harmany, "The sparse Poisson intensity reconstruction algorithms (SPIRAL) toolbox," [Online]. Available: http://drz.ac/code/spiraltap. Accessed on: Sep. 3, 2016.
[47] R. Gu, "Projected Nesterov's proximal-gradient algorithm source code," [Online]. Available: https://github.com/isucsp/pnpg. Accessed on: Jan. 18, 2017.
[48] J. A. Fessler, "Image reconstruction," 2009. [Online]. Available: http://web.eecs.umich.edu/~fessler/book/a-geom.pdf
[49] J. A. Fessler, "Image reconstruction toolbox," 2016. [Online]. Available: http://www.eecs.umich.edu/∼fessler/code. Accessed on: Aug. 23, 2016.
[50] A. Dogandžić, R. Gu, and K. Qiu, "Mask iterative hard thresholding algorithms for sparse image reconstruction of objects with known contour," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2011, pp. 2111–2116.
[51] S. Bonettini, I. Loris, F. Porta, and M. Prato, "Variable metric inexact line-search algorithm (VMILA)," [Online]. Available: http://www.oasis.unimore.it/site/home/software.html. Accessed on: Jan. 3, 2016.



Renliang Gu was born in Nantong, China. He received the B.S. degree in electrical engineering from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2009. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA.

His research interests include statistical signal processing, convex optimization, and X-ray computed tomography imaging.

Aleksandar Dogandžić (S'96–M'01–SM'06) received the Dipl. Ing. degree (*summa cum laude*) in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 1995, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of Illinois at Chicago, Chicago, IL, USA, in 1997 and 2001, respectively.

In August 2001, he joined the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA, where he is currently an Associate Professor. His research interests include statistical signal processing: theory and applications.

Dr. Dogandžić has served on editorial boards of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, THE IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, and the IEEE SIGNAL PROCESSING LETTERS. He has served as a General Co-chair of the Fourth and Fifth International Workshops on Computational Advances in Multi-Sensor Adaptive Processing in 2011 and 2013 and a Technical Co-chair of the 2014 and 2016 IEEE Sensor Array and Multichannel Workshops. He received the 2003 Young Author Best Paper Award and the 2004 Signal Processing Magazine Best Paper Award, both by the IEEE Signal Processing Society. In 2006, he received the CAREER Award by the National Science Foundation.