A Globally Linearly Convergent Method for Pointwise Quadratically Supportable Convex-Concave Saddle Point Problems

D. Russell Luke[*]and Ron Shefi[†]

February 28, 2017

Abstract

We study the *Proximal Alternating Predictor-Corrector* (PAPC) algorithm introduced recently by Drori, Sabach and Teboulle [8] to solve nonsmooth structured convex-concave saddle point problems consisting of the sum of a smooth convex function, a finite collection of nonsmooth convex functions and bilinear terms. We introduce the notion of pointwise quadratic supportability, which is a relaxation of a standard strong convexity assumption and allows us to show that the primal sequence is R-linearly convergent to an optimal solution and the primal-dual sequence is globally Q-linearly convergent. We illustrate the proposed method on total variation denoising problems and on locally adaptive estimation in signal/image deconvolution and denoising with multiresolution statistical constraints.

---

1   Introduction

We revisit the primal-dual first-order Proximal Alternating Predictor-Corrector (PAPC) splitting scheme introduced in [8] for solving a class of structured convex-concave saddle point problems involving a smooth function, a sum of finitely many bilinear terms and finitely many nonsmooth functions. This model covers a wide array of applications in signal/image processing and machine learning, see for instance [21, 19] and references therein.

This paper makes three contributions, two theoretical and one practical. In the first of the theoretical aspects, we introduce in Section 2 the notion of pointwise quadratic supportability, Definition 2.1, that is weaker than the typical assumption of strong convexity and allows one to treat common functions that, while not strongly convex, nevertheless possess desirable properties from a variational point of view. Pointwise quadratic supportability can be shown to be implied by pointwise strong monotonicity of the gradients (or more generally subdifferentials), which reveals a fundamental connection between this property and functions whose subdifferentials are *submonotone* as defined and studied in in [13]. A more detailed study of these connections, however, is not our focus here.

We are rather more narrowly focused on the second and third contributions of this paper, namely theoretical guarantees of global linear convergence of the PAPC algorithm studied in Section 3 and efficient implementations detailed in Section 4 for the problem of simultaneous denoising and deconvolution with multiresolution statistical constraints. Theorem 3.1 establishes global Q-linear convergence of the primal-dual iterates of the PAPC algorithm for convex problems with pointwise quadratically supportable objective functions and full rank linear mappings. A remarkable corollary of this result, Corollary 3.1, is that saddle point problems with this structure have unique solutions. In Corollary 3.2, we show global R-linear convergence of the primal sequence under the same assumptions. A sublinear iteration complexity in terms of the saddle point gap function values was established in [8] for the ergodic sequence. In our development, we show that global linear convergence of the *iterates* can be guaranteed under the assumption of pointwise quadratic supportability and full rank linear mappings. Based on these results, one can use stopping criteria that monitor only successive steps and provide an a posteriori error estimate on the distance to the set of exact solutions to the underlying problem.

We illustrate these results in Section 4 with two main applications, one conventional and one that is new research. The conventional example is of smoothed total variation (TV) denoising. The main motivation for us, however, is statistical multiscale image denoising/deconvolution following [10, 9] for fluorescence microscopic images (see also [2] for a

review of fluorescence microscopy techniques and statistical methods for them). The PAPC method is a full splitting approach where the gradient and the linear operators involved are called explicitly without inversion, while the simple nonsmooth functions are evaluated individually via their proximity operators. This allows for efficient implementations. The only other implementation we are aware of with similar convergence guarantees was reported in [1]. The numerical approach studied here can achieve results on the order of minutes on a small (32×32) image or less than an hour on a full 1024×1024 image. The approach presented in [1] required days for a 32×32 image and processing on the full-scale images was not feasible. As with [1], the convergence analysis presented here allows one to establish error bounds on solutions to the model problem which, in turn provide statistical guarantees on the numerical reconstruction [10]. This is demonstrated for image denoising/deconvolution of stimulated emission depletion (STED) images [11, 12]. An additional advantage of the PAPC algorithm and the analysis presented here is that a wider variety of regularizing objective functions can be incorporated easily.

Some well-known primal-dual decomposition methods are the Chambolle-Pock algorithm [5] and Alternating Direction of Multipliers (ADM). The method proposed by Chambolle-Pock – shown in [20] to be a equivalent to ADM with a weighted norm – requires the computation of proximal steps in each of the primal and the dual spaces. The algorithm has been shown to be linearly convergent when both functions are uniformly convex, but this assumption is far too stringent. Recently in [7], global linear convergence was shown under the assumptions of strict convexity and Lipschitz gradient on one of the two functions along with certain rank assumptions on the linear mapping. However, ADM-like methods do not easily extend to problems involving the sum of finitely many composite nonsmooth terms, see e.g., [6] where even the convergence becomes an issue. In [1] it was shown that the dual sequence generated by ADM – or equivalently the primal sequence of the Douglas-Rachford algorithm – must eventually achieve a linear rate of convergence from any starting point for piecewise linear-quadratic convex augmented Lagrangians whenever the saddle points are isolated.

1.1 Notation.

Our setting is the real vector space $\mathbb{R}^n$ with the norm generated from the inner product. The closed unit ball centered on the point $y \in \mathbb{R}^n$ is denoted by $\mathbb{B}(y)$. The domain of an extended real-valued function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is $\mathrm{dom}(\varphi) \equiv \{z \in \mathbb{R}^n : \varphi(z) < +\infty\}$. The Fenchel conjugate of φ is denoted by φ^* and is defined by $\varphi^*(u) = \sup_{z \in \mathbb{R}^n}\{\langle z, u \rangle - \varphi(z)\}$.

The set of symmetric $n \times n$ positive (semi)-definite matrices is denoted by $\mathbb{S}^n_{++}$ ($\mathbb{S}^n_+$). We use $M \succ 0$ ($M \succeq 0$) to denote a positive (semi)definite matrix. For any $z \in \mathbb{R}^n$ and any $M \in \mathbb{S}^n_+$, we denote the semi-norm $\|z\|^2_M := \langle z, Mz \rangle$. The operator norm is defined by $\|M\| = \max_{u \in \mathbb{R}^n}\{\|Mu\| : \|u\| = 1\}$ and coincides with the spectral radius of M whenever M is symmetric. If $A \neq 0$, $\sigma_{min}(A)$ denotes its smallest nonzero singular value. Let $\{z^k\}_{k \in \mathbb{N}}$ be a sequence that converges to z^*. We say the convergence is Q-linear if there exists $c \in (0,1)$ such that $\frac{\|z^{k+1}-z^*\|}{\|z^k-z^*\|} \leq c$ for all k; convergence is R-linear if there exists a sequence η_k such that $\|z^k - z^*\| \leq \eta_k$ and $\eta_k \to 0$ Q-linearly [16, Chapter 9].

We limit our discussion to proper (nowhere equal to $-\infty$ and finite at some point), lower semi-continuous (lsc), extended-valued (can take the value $+\infty$) functions. We will, in fact, limit our discussion to *convex* functions, but for our introduction of *pointwise quadratic supportability* in Definition 2.1 we formulate this with as much generality as possible to emphasize that convexity is not central to this key feature. By the *subdifferential* of a function φ, denoted $\partial\varphi$, we mean the collection of all *subgradients* that can be written as limits of sequences of *Fréchet subgradients* at nearby points; a vector v is a *(Fréchet) subgradient* of φ at y, written $v \in \widehat{\partial}\varphi(y)$, if

$$\liminf_{x \to y,\ x \neq y} \frac{\varphi(x) - \varphi(y) - \langle v, x - y \rangle}{\|x - y\|} \geq 0. \tag{1.1}$$

The functions of interest for us are *subdifferentially regular* on their domains, that is, the epigraphs of the functions are *Clarke regular* at points where they are finite [17, Definition 7.25]. For our purposes it suffices to note that, for a function φ that is subdifferentially regular at a point y, the subdifferential is nonempty and all subgradients are Fréchet subgradients, that is, $\partial\varphi(y) = \widehat{\partial}\varphi(y) \neq \emptyset$. Convex functions, in particular, are subdifferentially regular on their domains and the subdifferential has the particularly simple representation as the set of all vectors v where

$$\varphi(x) - \varphi(y) - \langle v, x - y \rangle \geq 0 \quad \forall x. \tag{1.2}$$

For $\varphi : \mathbb{R}^n \to (-\infty, \infty]$ a proper, lsc and convex function and for any $u \in \mathbb{R}^n$ and $M \in \mathbb{S}^n_{++}$, the proximal map associated with φ with respect to the weighted Euclidean norm is uniquely defined by:

$$\text{prox}^\varphi_M(u) = \underset{z}{\text{argmin}}\{\varphi(z) + \frac{1}{2}\|z - u\|^2_M : z \in \mathbb{R}^n\}.$$

When $M = c^{-1}I_n, c > 0$, we simply use the notation $\text{prox}^\varphi_c(u)$. We also recall the fundamental

Moreau proximal identity [14], that is, for any $z \in \mathbb{R}^n$

$$z = \text{prox}_M^\varphi(z) + M\text{prox}_{M^{-1}}^{\varphi^*}(M^{-1}(z)), \tag{1.3}$$

where M^{-1} is the inverse of $M \in \mathbb{S}_{++}^n$.

2 The saddle point model and the algorithm

This note focuses on the following *primal* problem:

$$p_* = \min_x \left\{ p(x) := f(x) + \sum_{i=1}^p g_i(A_i^T x) : x \in \mathbb{R}^n \right\}. \tag{$\mathcal{P}$}$$

The following blanket assumptions on the problem's data hold throughout:

Assumption 1.

(i) *The function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable with Lipschitz continuous gradient ∇f (constant L_f), that is for all $x, x' \in \mathbb{R}^n$, we have*

$$\|\nabla f(x) - \nabla f(x')\| \le L_f \|x - x'\|. \tag{2.1}$$

(ii) *$g_i : \mathbb{R}^{m_i} \to (-\infty, +\infty], i = 1, \ldots, p$ is proper, lsc, and convex.*

(iii) *The linear mappings $A_i : \mathbb{R}^{m_i} \to \mathbb{R}^n$, $i = 1, \ldots, p$ are full rank, that is, $\sigma_{min}^2(A_i) = \lambda_{min}(A_i^T A_i) > 0$.*

(iv) *The set of optimal solutions for problem $(\mathcal{P})$, denoted X^*, is nonempty.*

Assumption (ii) implies that the function defined by $g(y) := \sum_{i=1}^p g_i(y_i)$ with $y = (y_1, \ldots, y_p) \in \mathbb{R}^m$ for $m = \sum_{i=1}^p m_i$ is proper, lsc and convex. Assumption (iii) implies that the linear map $\mathcal{A} : \mathbb{R}^m \to \mathbb{R}^n$ given by $\mathcal{A}y = \sum_{i=1}^p A_i y_i$ is full rank. Assumption (iv) implies that the optimal value of the primal problem is finite.

The assumption of Lipschitz continuous gradients (i), while standard in the literature involving complexity bounds, is very strong indeed. A more satisfying theory would involve only functions with Lipschitz continuous gradients on bounded domains. But this would involve a reinvention of much of the theory. The main issue is boundedness of the iterates,

which is not a precursor to the convergence results of [8] upon which we build. Four our purposes, Lipschitz continuity suffices, and leaves a short path to the main result. We leave the stronger results as an open challenge.

We reformulate the problem $(\mathcal{P})$ as a convex-concave saddle point problem, and then apply the primal-dual algorithm in [8] to find a saddle point solution. The *primal-dual* problem associated to $(\mathcal{P})$ consists of finding a saddle point of the Lagrangian:

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} \left\{ K(x, y) := f(x) + \langle x, \mathcal{A}y \rangle - g^*(y) \right\}. \qquad (\mathcal{M})$$

Assumption 1(iv) guarantees that the convex-concave function $K(\cdot, \cdot)$ has a saddle point, that is, there exists $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$K(\hat{x}, y) \leq K(\hat{x}, \hat{y}) \leq K(x, \hat{y}) \qquad \forall x \in \mathbb{R}^n, y \in \mathbb{R}^m.$$

The existence of a saddle point corresponds to zero duality gap for the induced optimization problems

$$p(x) = \sup_y \{K(x, y) : y \in \mathbb{R}^m\} \qquad q(y) = \inf_x \{K(x, y) : x \in \mathbb{R}^n\}.$$

By weak duality, we have $\inf_{x \in \mathbb{R}^n} p(x) \geq \sup_{y \in \mathbb{R}^m} q(y)$. In addition, under standard constraint qualifications (e.g., [3, Chapter 5] or [4, Theorem 2.3.4]), $(\hat{x}, \hat{y})$ is a saddle point of K if and only if $\hat{x}$ is an optimal solution of the primal problem $(\mathcal{P})$, $\hat{y}$ is an optimal solution of the dual problem to $(\mathcal{P})$.

Constraints are included in the model through the extended-valued functions g_i, which need not be smooth. The properties of the function f are crucial to the success of the algorithm. In addition to Assumption 1 we will assume that f is *pointwise quadratically supportable*.

Definition 2.1 (pointwise quadratically supportable mappings)**.**

(i) *A proper, extended-valued function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be* pointwise quadratically supportable at y *if it is subdifferentially regular there and there exists a neighborhood V of y and a constant $\mu > 0$ such that*

$$(\forall v \in \partial \varphi(y)) \quad \varphi(x) \geq \varphi(y) + \langle v, x - y \rangle + \frac{\mu}{2} \|x - y\|^2, \quad \forall x \in V. \qquad (2.2)$$

If for each bounded neighborhood V of y there exists a constant $\mu > 0$ such that (2.2) holds, then the function φ is said to be pointwise quadratically supportable at y on

6

bounded sets. *If (2.2) holds with one and the same constant $\mu > 0$ on all neighborhoods V, then φ is said to be* uniformly *pointwise quadratically supportable at y.*

(ii) *A proper, extended-valued function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be* strongly coercive *at y if it is subdifferentially regular on V and there exists a neighborhood V of y and a constant $\mu > 0$ such that*

$$(\forall v \in \partial\varphi(z)) \quad \varphi(x) \geq \varphi(z) + \langle v, x - z \rangle + \frac{\mu}{2}\|x - z\|^2, \quad \forall x, z \in V. \qquad (2.3)$$

If for each bounded neighborhood V of y there exists a constant $\mu > 0$ such that (2.3) holds, then the function φ is said to be strongly coercive at y on bounded sets. *If (2.3) holds with one and the same constant $\mu > 0$ on all neighborhoods V, then φ is said to be* uniformly *strongly coercive at y.*

Clearly strong coercivity implies pointwise quadratic supportability, but the reverse implication does not hold, as the next example shows.

Example 2.1. *The Huber function (see (4.3) in Section 4) common in robust regression is an example of a smooth convex function that does not satisfy (2.2) at all points (namely where it is linear), but does satisfy this inequality at its minimum. A slight modification of the Huber function is*

$$\phi_\alpha(t) = \begin{cases} \frac{(t+\epsilon)^2 - \epsilon^2}{2\alpha} & \text{if } 0 \leq t \leq \alpha - \epsilon \\ \frac{(t-\epsilon)^2 - \epsilon^2}{2\alpha} & \text{if } -\alpha + \epsilon \leq t \leq 0 \\ |t| + \left(\epsilon - \frac{\epsilon^2 + \alpha^2}{2\alpha}\right) & \text{if } |t| > \alpha - \epsilon. \end{cases} \qquad (2.4)$$

This function is convex and pointwise quadratically supportable on bounded sets at $t = 0$ where $\partial\phi_\alpha(0) = [-\epsilon/\alpha, \epsilon/\alpha]$. This function is not, however, strongly convex due to the linear portion for $|t| > \alpha - \epsilon$.

For convex functions, pointwise quadratic supportability is *equivalent* to pointwise quadratic supportability on bounded sets, as the next proposition establishes.

Proposition 2.1. *Let $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper extended-valued and convex on $\mathbb{R}^n$. The following are equivalent.*

(i) *φ is pointwise quadratically supportable at y;*

(ii) *φ is pointwise quadratically supportable at y on bounded sets.*

7

Proof. That pointwise quadratic supportability at y on bounded sets implies pointwise quadratic supportability at y is clear. For the converse implication, fix $R > 0$ and choose any $z \in R\mathbb{B}(y)$. Let V be the neighborhood of y on which (2.2) holds with constant μ. If $z \in V$, then inequality (2.2) is trivially satisfied. Suppose, then, that $z \notin V$. Since V is a neighborhood of y, there exists a $\delta > 0$ and a point $x \in \delta\mathbb{B}(y) \subset V$ such that $x = \tau z + (1-\tau)y$ for $\tau \in (0,1)$ and $\|x - y\| = \delta$. Then

$$z - y = \frac{1}{1-\tau}(z - x) = \frac{1}{\tau}(x - y)$$

and

$$\tau = \frac{\delta}{\|z - y\|} \geq \frac{\delta}{R}.$$

By the characterization of the convex subdifferential we have

$$\varphi(z) \geq \varphi(y) + \langle v, z - y \rangle \ \forall \ v \in \partial\varphi(y).$$

Adding $\varphi(x) - \varphi(x)$ to this and using (2.2) yields

$$
\begin{aligned}
(\forall \ v \in \partial\varphi(y)) \qquad \varphi(z) &\geq \varphi(y) + \left(\varphi(y) + \langle v, x - y \rangle + \frac{\mu}{2}\|x - y\|^2\right) - \varphi(x) + \langle v, z - y \rangle \\
&= 2\varphi(y) + (1+\tau)\langle v, z - y \rangle + \frac{\mu\tau^2}{2}\|z - y\|^2 - \varphi(x) \\
&\geq 2\varphi(y) + (1+\tau)\langle v, z - y \rangle + \frac{\mu\tau^2}{2}\|z - y\|^2 - (\tau\varphi(z) + (1-\tau)\varphi(y)) \\
&= (1+\tau)\varphi(y) - \tau\varphi(z) + (1+\tau)\langle v, z - y \rangle + \frac{\mu\tau^2}{2}\|z - y\|^2.
\end{aligned}
$$

Rearranging the terms and simplifying yields

$$(\forall \ v \in \partial\varphi(y)) \qquad \varphi(z) \geq \varphi(y) + \langle v, z - y \rangle + \frac{\mu\tau^2}{2(1+\tau)}\|z - y\|^2 \geq \varphi(y) + \langle v, z - y \rangle + \frac{\mu\delta^2}{4R^2}\|z - y\|^2.$$

Since R and z are arbitrary, this completes the proof. □

It is worthwhile contrasting the above property to the assumption of *strong convexity* that is common in the literature. Analogous to Definition 2.1(i), we generalize this notion to *pointwise strong convexity*, which is new.

Definition 2.2 ((pointwise) strongly convex functions).

(i) *A function* $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is said to be* pointwise strongly convex at y *if there exists a convex neighborhood* V *of* y *and a constant* $\mu > 0$ *such that,*

$$(\forall \tau \in (0,1)) \quad \varphi(\tau x + (1-\tau)y) \le \tau \varphi(x) + (1-\tau)\varphi(y) - \frac{1}{2}\mu\tau(1-\tau)\|x-y\|^2, \quad \forall x \in V. \tag{2.5}$$

If for each bounded convex neighborhood V *there exists a constant* $\mu > 0$ *such that* (2.5) *holds, then* φ *is said to be* pointwise strongly convex at y on bounded sets. *If there exists a single constant* $\mu > 0$ *such that* (2.5) *holds on all convex neighborhoods* V, *then* φ *is said to be* uniformly *pointwise strongly convex at* y.

(ii) *A function* $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is said to be* strongly convex at y *if there exists a convex neighborhood* V *of* y *and a constant* $\mu > 0$ *such that,*

$$(\forall \tau \in (0,1)) \quad \varphi(\tau x + (1-\tau)z) \le \tau \varphi(x) + (1-\tau)\varphi(z) - \frac{1}{2}\mu\tau(1-\tau)\|x-z\|^2, \quad \forall x, z \in V. \tag{2.6}$$

If for each bounded convex neighborhood V *there exists a constant* $\mu > 0$ *such that* (2.6) *holds, then* φ *is said to be* strongly convex at y on bounded sets. *If there exists a single constant* $\mu > 0$ *such that* (2.6) *holds on all convex neighborhoods* V, *then* φ *is said to be* uniformly *strongly convex at* y.

Again, it is clear from the definition that strong convexity implies pointwise strong convexity, but the converse need not, in general, be true. Indeed, it is well known in the smooth case (φ continuously differentiable) that strong convexity defined by (2.6) is *equivalent* to strong coercivity defined by (2.3), with the same constants on the same neighborhoods. The situation is very different for the pointwise definition, even for smooth functions as the next example demonstrates.

Example 2.2 (pointwise quadratic supportability does not imply convexity)**.** *The function* $\varphi(y) \equiv 1 - e^{-y^2}$ *is pointwise quadratically supportable on bounded sets at* $y = 0$, *but it is neither convex nor pointwise strongly convex on bounded sets. For that matter, it is not even coercive.*

Another notion common in the literature is strong monotonicity. Again, for a smooth convex function, strong convexity is equivalent to strong monotonicity of its gradient. The next result shows that pointwise strong convexity and pointwise strong monotonicity for differentiable functions imply pointwise quadratic supportability. Together with the above example, this shows that pointwise strong coercivity is the weakest of the three notions.

Proposition 2.2 (pointwise quadratically supportable/convex/monotone differentiable functions). *Let $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be continuously differentiable at y.*

(i) *If φ is pointwise strongly convex at y with constant μ on the convex neighborhood V (that is, φ satisfies (2.5)), then φ is pointwise quadratically supportable at y with constant μ on V (that is, φ satisfies (2.2)).*

(ii) *If there exists a constant μ and a convex neighborhood V of y such that*

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x - y \rangle \geq \mu \|x - y\|^2 \quad \forall x \in V, \tag{2.7}$$

then φ satisfies (2.2) with the same constant on some neighborhood $V' \subset V$.

Proof. To see (i), rearrange the inequality (2.5), divide through by τ and take the limit as $\tau \to 0$.

To see (ii) choose any $x \in V$. By the Mean Value Theorem there is a $\lambda \in (0, 1)$ such that, for $z = \lambda x + (1 - \lambda)y$ (which is in V since this is convex),

$$
\begin{aligned}
\varphi(x) - \varphi(y) &= \langle \nabla\varphi(z), x - y \rangle \\
&= \langle \nabla\varphi(y), x - y \rangle + \langle \nabla\varphi(z) - \nabla\varphi(y), x - y \rangle \\
&= \langle \nabla\varphi(y), x - y \rangle + \frac{1}{\lambda} \langle \nabla\varphi(z) - \nabla\varphi(y), z - y \rangle \\
&\geq \langle \nabla\varphi(y), x - y \rangle + \frac{\mu}{\lambda} \|z - y\|^2 \\
&\geq \langle \nabla\varphi(y), x - y \rangle + \lambda\mu \|x - y\|^2.
\end{aligned}
$$

By continuity of $\nabla\varphi$, $\lambda \to 1$ as $x \to y$, hence, for all x close enough to y, that is in some neighborhood $V' \subset V$, $\lambda > 1/2$ and

$$\varphi(x) \geq \varphi(y) + \langle \nabla\varphi(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$$

as claimed. $\square$

We leave a complete development of pointwise coercivity and related objects to future research. For our concrete application, namely linear image denoising and deconvolution, we will assume the following throughout.

> **Assumption 2.** *The function $f : \mathbb{R}^n \to \mathbb{R}$ of Problem $\mathcal{P}$ is pointwise quadratically supportable at each $\hat{x} \in X^*$, that is, there exists a $\mu > 0$ and a convex neighborhood V of $\hat{x}$ such that*
>
> $$f(x) \geq f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle + \tfrac{1}{2}\mu \|x - \hat{x}\|^2 \quad \forall x \in V. \tag{2.8}$$

Note that condition (ii) of Proposition 2.2 puts a lower bound on the constant of pointwise Lipschitz continuity for ∇f, namely μ. Also, as a consequence of Proposition 2.1 any convex function satisfying Assumption 2 is pointwise quadratically supportable at all points in the solution set X^* on all bounded convex neighborhoods of these points. This property, together with Assumption 1 will yield global linear convergence for our proposed algorithm and, as a corollary, uniqueness of saddlepoints.

2.1 The Algorithm

The algorithm we revisit is the primal-dual (PAPC) algorithm proposed in [8] for solving $(\mathcal{M})$. It consists of a predictor-corrector gradient step for handling the smooth part of K and a proximal step for handling the nonsmooth part.

> **Proximal Alternating Predictor-Corrector (PAPC) for solving $(\mathcal{M})$**
>
> **Initialization:** Let $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m$, and choose the parameters τ and σ to satisfy
>
> $$\tau \in \left(0, \frac{1}{L_f}\right), \quad 0 < \tau\sigma \leq \tfrac{1}{\|\mathcal{A}^T \mathcal{A}\|}. \tag{2.9a}$$
>
> **Main Iteration:** for $k = 1, 2, \ldots$ update x^k, y^k as follows:
>
> $$p^k = x^{k-1} - \tau(\nabla f(x^{k-1}) + \mathcal{A}y^{k-1}); \tag{2.9b}$$
>
> for $i = 1, \ldots, p$,
> $$y_i^k = \operatorname*{argmax}_{y_i \in \mathbb{R}^{m_i}} \{ \langle A_i^T p^k, y_i \rangle - g_i^*(y_i) - (1/2\sigma)\|y_i - y_i^{k-1}\|^2 \} \equiv \operatorname{prox}_\sigma^{g_i^*}(y_i^{k-1} + \sigma A_i^T p^k);$$
> $$\tag{2.9c}$$
>
> $$x^k = x^{k-1} - \tau(\nabla f(x^{k-1}) + \mathcal{A}y^k). \tag{2.9d}$$

At each iteration the algorithm utilizes one gradient and full proximal map evaluation on the given nonsmooth function, assumed to be easy to compute. For our purposes, we suppose these can be evaluated exactly, though with finite precision arithmetic this is clearly

not realistic. A version of this theory that takes inexactness into account is left to future research.

The step in the dual space (2.9c) can be written in a short form using the following notation. Since $g(y) := g(y_1, \ldots, y_p) = \sum_{i=1}^{p} g_i(y_i)$, then the convex conjugate of a separable sum of functions gives $g^*(y) := \sum_{i=1}^{p} g_i^*(y_i)$ and the definition of the matrix $S = \sigma^{-1} I_m$ we immediately get that for any point $\zeta_i \in \mathbb{R}^{m_i}$, $i = 1, \ldots, m$,

$$\text{prox}_S^{g^*}(\zeta) = (\text{prox}_\sigma^{g_1^*}(\zeta_1), \text{prox}_\sigma^{g_2^*}(\zeta_2), \ldots, \text{prox}_\sigma^{g_p^*}(\zeta_p)).$$

Thus the step (2.9c) can be written in vector notation by $y^k = \text{prox}_S^g(y^{k-1} + \sigma \mathcal{A}^T p^k)$. It is possible to use different proximal step constants σ_i, $i = 1 \ldots, p$, see details in [8]. The choice $\sigma_i = \sigma$ for $i = 1, \ldots, p$ is purely for simplicity of exposition.

In [8] the parameter choices for obtaining the iteration complexity with respect to the ergodic sequence was given and the convergence of the sequence $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ to a saddle point solution was established. As we will use this later, we quote the main result here.

Proposition 2.3. *[8, Corollary 3.2] Let $\{(p^k, y^k, x^k)\}_{k \in \mathbb{N}}$ be the sequence generated by the PAPC algorithm. If Assumptions 1(i)(ii)(iv) are satisfied then the sequence $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ converges to a saddle point $(\bar{x}, \bar{y})$ of $K(\cdot, \cdot)$.*

The next intermediate result establishes pointwise quadratic supportability (Definition 2.1) on bounded sets at all saddle points under Assumptions 1 and 2.

Proposition 2.4. *Let $\{(p^k, y^k, x^k)\}_{k \in \mathbb{N}}$ be the sequence generated by the PAPC algorithm. If Assumptions 1 and 2 are satisfied, then for any primal solution $\hat{x}$ to the saddle point problem $(\mathcal{M})$, there exists a $\mu > 0$ such that*

$$f(x^k) \geq f(\hat{x}) + \langle \nabla f(\hat{x}), x^k - \hat{x} \rangle + \tfrac{1}{2}\mu \|x^k - \hat{x}\|^2 \quad \forall k. \tag{2.10}$$

Proof. Let $\hat{x}$ be a primal solution to the saddle point problem. By Proposition 2.3 the sequence $\{x^k\}_{k \in \mathbb{N}}$ is bounded and indeed converges to a primal solution $\tilde{x}$, not necessarily the same point as $\hat{x}$. A function f satisfying Assumption 2 is pointwise quadratically supportable at any primal saddle point solution to $(\mathcal{M})$, and hence by Proposition 2.1, f is pointwise quadratically supportable at $\hat{x}$ on bounded convex sets. In other words, there exists a $\mu > 0$ and a ball $R\mathbb{B}(\hat{x})$ containing the sequence $\{x^k\}_{k \in \mathbb{N}}$ on its interior such that f satisfies (2.10), as claimed. □

The constant μ in Proposition 2.4 depends on the choice of (x^0, y^0) and so depends implicitly on the distance of the initial guess to the point in the set of saddle point solutions.

3 Convergence of the iterates

In this section we show that, under Assumptions 1 and 2, the *primal* sequence generated by the PAPC algorithm converges R-linearly to a saddle point. We recall the following simple observations (see [8]).

Proposition 3.1. *Let* $(\hat{x}, \hat{y})$ *be a saddle point of* K, *and let* $\{(p^k, y^k, x^k)\}_{k \in \mathbb{N}}$ *be the sequence generated by PAPC. Then, for all* $k \geq 1$,

$$\nabla f(x^{k-1}) - \nabla f(\hat{x}) + \mathcal{A}(y^k - \hat{y}) + \frac{1}{\tau}(x^k - x^{k-1}) = 0. \tag{3.1}$$

Proof. This follows immediately from the update rule (2.9d) together with the optimality condition for $(\mathcal{M})$ $\nabla f(\hat{x}) + \mathcal{A}\hat{y} = 0$ where $(\hat{x}, \hat{y})$ is a saddle point solution to $(\mathcal{M})$. $\qquad\square$

The next lemma uses the following shorthand notation:

$$G := \sigma^{-1} I_m - \tau \mathcal{A}^T \mathcal{A}. \tag{3.2}$$

Note that for the choice of τ given in (2.9a), $G \succeq 0$.

Lemma 3.1. *Let* $\{(p^k, x^k, y^k)\}_{k \in \mathbb{N}}$ *be the sequence generated by the PAPC algorithm. Then for every* $k \in \mathbb{N}$ *and for any* $x \in \mathbb{R}^n$ *and* $y \in \mathbb{R}^m$

$$\begin{aligned}
K(x^k, y) - K(x, y^k) \leq{} & \frac{1}{2} \left(\|y^{k-1} - y\|_G^2 - \|y^k - y\|_G^2 - \|y^{k-1} - y^k\|_G^2 \right) \\
& + \frac{1}{2\tau} \left(\|x^{k-1} - x\|^2 - \|x^k - x\|^2 \right) - \frac{1}{2} \left(\frac{1}{\tau} - L_f \right) \|x^k - x^{k-1}\|^2.
\end{aligned} \tag{3.3}$$

Proof.
See Lemma [8, Lemma 3.1]. $\qquad\square$

The next lemma is key to the global linear convergence results presented in the next subsection.

Lemma 3.2. *Let $(\hat{x}, \hat{y})$ be a saddle point solution for $K(\cdot, \cdot)$ and let $(p^k, x^k, y^k)_{k \in \mathbb{N}}$ be the sequence generated by the PAPC algorithm. If Assumptions 1 and 2 are satisfied, then there exists a $\mu > 0$ such that for every $k \in \mathbb{N}$*

$$
\begin{aligned}
\frac{\mu}{2} \|x^k - \hat{x}\|^2 \leq{}& \frac{1}{2} \left(\|y^{k-1} - \hat{y}\|_G^2 - \|y^k - \hat{y}\|_G^2 - \|y^{k-1} - y^k\|_G^2 \right) \\
&+ \frac{1}{2\tau} \left(\|x^{k-1} - \hat{x}\|^2 - \|x^k - \hat{x}\|^2 \right) - \frac{1}{2} \left(\frac{1}{\tau} - L_f \right) \|x^k - x^{k-1}\|^2.
\end{aligned}
\tag{3.4}
$$

Proof. Fix $(\hat{x}, \hat{y})$, the saddle point solution to $(\mathcal{M})$. By (3.3) of Lemma 3.1, we have

$$
\begin{aligned}
K(x^k, \hat{y}) - K(\hat{x}, y^k) &= K(x^k, \hat{y}) - K(\hat{x}, \hat{y}) + K(\hat{x}, \hat{y}) - K(\hat{x}, y^k) \\
&\leq \frac{1}{2} \left(\|y^{k-1} - \hat{y}\|_G^2 - \|y^k - \hat{y}\|_G^2 - \|y^{k-1} - y^k\|_G^2 \right) \\
&\quad + \frac{1}{2\tau} \left(\|x^{k-1} - \hat{x}\|^2 - \|x^k - \hat{x}\|^2 \right) - \frac{1}{2} \left(\frac{1}{\tau} - L_f \right) \|x^k - x^{k-1}\|^2 \tag{3.5}
\end{aligned}
$$

Recall that for the parameter values specified in (2.9a) $G \succeq 0$. By the saddle point inequality $K(\hat{x}, \hat{y}) - K(\hat{x}, y^k) \geq 0$, and by Proposition 2.4 there is a $\mu > 0$ such that

$$
K(x^k, \hat{y}) - K(\hat{x}, \hat{y}) = f(x^k) - f(\hat{x}) + \langle x^k - \hat{x}, \mathcal{A}\hat{y} \rangle = f(x^k) - f(\hat{x}) - \langle x^k - \hat{x}, \nabla f(\hat{x}) \rangle \geq \frac{\mu}{2} \|x^k - \hat{x}\|^2,
$$

where in the last inequality we have used optimality condition of problem $(\mathcal{M})$, namely $\nabla f(\hat{x}) + \mathcal{A}\hat{y} = 0$, and the result follows. $\qquad\square$

Note that the saddle point in Lemma 3.2 need not be the limit point of the sequence generated by PAPC.

3.1   Main Results - R-Linear Rate of the Primal Sequence

Convergence of the primal-dual sequence becomes more transparent using G in (3.2) to define a weighted norm on the primal-dual product space.

$$
u = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad H = \begin{pmatrix} \tau^{-1} I_n & 0 \\ 0 & G \end{pmatrix}
\tag{3.6}
$$

where by the assumptions on the choice of τ given in (2.9a), $G \succeq 0$. We can then define an associated norm using the positive-semidefinite matrix H, $\|u\|_H^2 := \frac{1}{\tau}\|x\|^2 + \|y\|_G^2$. The main

inequality in Lemma 3.2 written in terms of this norm then becomes

$$\left(\frac{1}{\tau} - L_f\right) \|x^k - x^{k-1}\|^2 + \|y^{k-1} - y^k\|_G^2 + \mu\|x^k - \hat{x}\|^2 \le \|u^{k-1} - \hat{u}\|_H^2 - \|u^k - \hat{u}\|_H^2. \quad (3.7)$$

In order to establish the Q-linear (global) convergence of the associated sequence $\{u^k\}_{k\in\mathbb{N}}$ with respect to the H-norm, we show that there exists a positive δ such that

$$\forall k \in \mathbb{N}: \qquad (1+\delta)\|u^k - \hat{u}\|_H^2 \le \|u^{k-1} - \hat{u}\|_H^2. \qquad (3.8)$$

This will depend implicitly on μ, the constant of pointwise strong convexity of f, which depends implicitly on the distance of the initial guess to the solution set X^*. In Theorem 3.1 we will show there exists a positive δ such that for all $k \ge 1$,

$$\delta\|u^k - \hat{u}\|_H^2 \le \left(\frac{1}{\tau} - L_f\right) \|x^k - x^{k-1}\|^2 + \mu\|x^k - \hat{x}\|^2, \qquad (3.9)$$

Since adding (3.9) to (3.7) (ignoring the term $\|y^{k-1} - y^k\|_G^2$ which is nonnegative) yields (3.8), we obtain the Q-linear rate for the sequence $\{u^k\}_{k\in\mathbb{N}}$ with respect to the H-norm.

Theorem 3.1. *Let $\{(p^k, x^k, y^k)\}_{k\in\mathbb{N}}$ be the sequence generated by the PAPC algorithm, and let $(\hat{x}, \hat{y})$ be any saddle point solution for $K(\cdot, \cdot)$. If Assumptions 1 and 2 are satisfied, then, for any $\alpha > 1$ and for all $k \ge 1$, the sequence $\{u^k\}_{k\in\mathbb{N}}$ satisfies*

$$\|u^k - \hat{u}\|_H^2 \le \frac{1}{1+\delta}\|u^{k-1} - \hat{u}\|_H^2, \qquad (3.10)$$

where
$$\delta = \min\left\{ \frac{(\alpha-1)\tau\sigma(1-\tau L_f)\lambda_{min}(\mathcal{A}^T\mathcal{A})}{\alpha}, \frac{\mu\tau\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})}{\alpha\tau L_f^2 + \sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})} \right\} \qquad (3.11)$$

is positive and $\mu > 0$ is the constant of pointwise quadratic supportability of f at $\hat{x}$ depending on the distance of the initial guess to the point $(\hat{x}, \hat{y})$ in the solution set X^*. In particular, $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ is Q-linearly convergent with respect to the H-norm to a saddle-point solution.

Proof. Let $(\hat{x}, \hat{y})$ be any saddle point solution for $K(\cdot, \cdot)$. Using (3.1) in Proposition 3.1

$$\frac{1}{\tau}(x^k - x^{k-1}) = \nabla f(\hat{x}) - \nabla f(x^{k-1}) + \mathcal{A}(\hat{y} - y^k),$$

and by adding and subtracting $\nabla f(x^k)$ and rearranging, we have

$$\frac{1}{\tau}(x^k - x^{k-1}) - \left(\nabla f(x^k) - \nabla f(x^{k-1})\right) = \nabla f(\hat{x}) - \nabla f(x^k) + \mathcal{A}(\hat{y} - y^k). \tag{3.12}$$

Applying the inequality $\|a + b\|^2 \geq (1 - \alpha)\|a\|^2 + (1 - \frac{1}{\alpha})\|b\|^2 \quad \forall \alpha > 0$ to (3.12) with $a := \nabla f(\hat{x}) - \nabla f(x^k)$, and $b := \mathcal{A}(\hat{y} - y^k)$, we have

$$\left\| \frac{1}{\tau}\left(x^k - x^{k-1}\right) - \left(\nabla f(x^k) - \nabla f(x^{k-1})\right) \right\|^2$$
$$\geq (1 - \alpha)\|\nabla f(\hat{x}) - \nabla f(x^k)\|^2 + \left(1 - \frac{1}{\alpha}\right)\|\mathcal{A}(\hat{y} - y^k)\|^2. \tag{3.13}$$

Next we have

$$\left\| \frac{1}{\tau}\left(x^k - x^{k-1}\right) - \left(\nabla f(x^k) - \nabla f(x^{k-1})\right) \right\|^2$$
$$= \frac{1}{\tau^2}\left\|x^k - x^{k-1}\right\|^2 - \frac{2}{\tau}\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle + \|\nabla f(x^k) - \nabla f(x^{k-1})\|^2$$
$$\leq \frac{1}{\tau^2}\left\|x^k - x^{k-1}\right\|^2 + \left(1 - \frac{2}{\tau L_f}\right)\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2 \tag{3.14a}$$
$$\leq \frac{1}{\tau^2}\|x^k - x^{k-1}\|^2, \tag{3.14b}$$

where in inequality (3.14a) we used $\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle \geq \frac{1}{L_f}\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2$ (cf. [15, Thm. 2.1.5]), and in inequality (3.14b) we used the fact that $\tau L_f < 1$. Adding (3.13)-(3.14) we get

$$\frac{1}{\tau^2}\|x^k - x^{k-1}\|^2 + (\alpha - 1)\|\nabla f(\hat{x}) - \nabla f(x^k)\|^2 \geq \left(1 - \frac{1}{\alpha}\right)\|\mathcal{A}(\hat{y} - y^k)\|^2.$$

For any $\alpha > 1$ (so that $1 - \frac{1}{\alpha} > 0$), we can bound from below the right-hand side of the later inequality by Assumption 1(iii) and bound from above the left-hand side by Assumption 1(i) to yield

$$\frac{1}{\tau^2}\|x^k - x^{k-1}\|^2 + (\alpha - 1)L_f^2\|\hat{x} - x^k\|^2 \geq \left(1 - \frac{1}{\alpha}\right)\lambda_{min}(\mathcal{A}^T\mathcal{A})\|\hat{y} - y^k\|^2.$$

Using the semi-norm notation to write $\|\hat{y} - y^k\|^2 = \sigma\|\hat{y} - y^k\|^2_{\sigma^{-1}I_m}$ yields

$$\frac{1}{\tau^2}\|x^k - x^{k-1}\|^2 + (\alpha - 1)L_f^2\|\hat{x} - x^k\|^2 \geq \left(1 - \frac{1}{\alpha}\right)\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})\|\hat{y} - y^k\|^2_{\sigma^{-1}I_m}.$$

16

Subtracting from both sides of the inequality the term

$$\left(1 - \frac{1}{\alpha}\right)\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})\|\hat{y} - y^k\|^2_{\tau\mathcal{A}^T\mathcal{A}}$$

(which is positive since $\alpha > 1$, $\tau > 0$ and $\mathcal{A}^T\mathcal{A} \in \mathbb{S}^m_{++}$) together with the notation of $\|\cdot\|_G$ defined in (3.2), yields

$$\frac{1}{\tau^2}\|x^k - x^{k-1}\|^2 + (\alpha - 1)L_f^2\|\hat{x} - x^k\|^2 \geq \left(1 - \frac{1}{\alpha}\right)\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})\|\hat{y} - y^k\|_G^2.$$

After multiplication of both sides by $\frac{\alpha}{(\alpha-1)\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})}$, we get

$$\left(\frac{\alpha}{\alpha - 1}\right)\frac{1}{\sigma\tau^2\lambda_{min}(\mathcal{A}^T\mathcal{A})}\|x^k - x^{k-1}\|^2 + \frac{\alpha L_f^2}{\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})}\|\hat{x} - x^k\|^2 \geq \|\hat{y} - y^k\|_G^2.$$

Finally, adding to both sides of the inequality $\frac{1}{\tau}\|x^k - \hat{x}\|^2$, and the definition of the associated norm $\|\hat{u} - u^k\|_H^2 := \frac{1}{\tau}\|\hat{x} - x^k\|^2 + \|\hat{y} - y^k\|_G^2$, we have

$$\left(\frac{\alpha}{\alpha - 1}\right)\frac{1}{\sigma\tau^2\lambda_{min}(\mathcal{A}^T\mathcal{A})}\|x^k - x^{k-1}\|^2 + \frac{\alpha\tau L_f^2 + \sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})}{\tau\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})}\|\hat{x} - x^k\|^2 \geq \|\hat{u} - u^k\|_H^2. \quad (3.15)$$

Now, for $\mu > 0$ satisfying (3.4) in Lemma 3.2, we choose δ so that

$$\frac{1}{\tau} - L_f \geq \left[\left(\frac{\alpha}{\alpha - 1}\right)\frac{1}{\sigma\tau^2\lambda_{min}(\mathcal{A}^T\mathcal{A})}\right]\delta, \qquad \mu \geq \left[\frac{\alpha\tau L_f^2 + \sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})}{\tau\sigma\lambda_{min}(\mathcal{A}^T\mathcal{A})}\right]\delta. \quad (3.16)$$

This establishes the choice of δ in (3.11). Moreover, δ is positive since $\tau \in (0, 1/L_f)$ and $\alpha > 1$. Multiplying the inequality (3.15) by this δ, we have

$$\left(\frac{1}{\tau} - L_f\right)\|x^k - x^{k-1}\|^2 + \mu\|x^k - \hat{x}\|^2 \geq \delta\|u^k - \hat{u}\|_H^2.$$

Adding the above inequality to (3.4) in Lemma 3.2 yields (3.10). Since by Proposition 2.3 sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ converges to a saddle point $\bar{u} \equiv (\bar{x}, \bar{y})$ and the choice of saddle point $(\hat{x}, \hat{y})$ in the argument above was arbitrary, we have, for δ satisfying (3.11),

$$\|u^k - \bar{u}\|_H^2 \leq \frac{1}{1 + \delta}\|u^{k-1} - \bar{u}\|_H^2 \qquad (3.17)$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

In fact, the above theorem about convergence of the PAPC algorithm actually proves a much more striking fact that is independent of the algorithm.

Corollary 3.1 (unique saddlepoint). *If Assumptions 1 and 2 are satisfied then the solution set X^* is a singleton.*

Proof. Inequality (3.10) holds for all k for any point $\hat{u} \in X^*$, thus $u^k \to \hat{u}$ as $k \to \infty$ for all $\hat{u} \in X^*$, which can only happen if $X^* = \{\bar{u}\}$, the limit of the sequence $\{u^k\}_{k\in\mathbb{N}}$. $\square$

We also immediately obtain the following corollary, which establishes the expected R-linear iteration complexity for the primal sequence of PAPC for obtaining ϵ-optimal solution. A sublinear complexity of $O(1/\epsilon)$ in terms of saddle point gap function values was shown in [8, Corollary 3.1] for the ergodic sequence. In the next corollary an R-linear rate can be guaranteed under the assumption of pointwise quadratic supportability and the full rank assumption on the linear mappings (Assumption 1(iii)).

Corollary 3.2. *Suppose Assumptions 1 and 2 are satisfied and let $\bar{u} = (\bar{x}, \bar{y})$ be the limit point of the sequence generated by the PAPC algorithm. In order obtain*

$$\|x^k - \bar{x}\| \leq \epsilon \qquad (resp. \quad \|y^k - \bar{y}\|_G \leq \epsilon), \tag{3.18}$$

it suffices to compute k iterations, with

$$k \geq \frac{2\log\left(\frac{C}{L_f\epsilon}\right)}{\delta} \qquad \left(resp. \quad k \geq \frac{2\log\left(\frac{C}{\epsilon}\right)}{\delta}\right), \tag{3.19}$$

where $C = \|u^0 - \bar{u}\|_H = \left(\frac{1}{\tau}\|x^0 - \bar{x}\|^2 + \|y^0 - \bar{y}\|_G^2\right)^{1/2}$, and δ is given in (3.11).

Proof. We have by Theorem 3.1 that $\|u^k - \bar{u}\| \leq C\omega^{k/2}$, where

$$C = \|u^0 - \bar{u}\|_H = \left(\frac{1}{\tau}\|x^0 - \bar{x}\|^2 + \|y^0 - \bar{y}\|_G^2\right)^{1/2} \quad \text{and} \quad \omega = (1 + \delta)^{-1}.$$

Hence,
$$\|x^k - \bar{x}\| \leq \tau C\omega^{k/2} < \frac{C}{L_f}\omega^{k/2}, \quad \|y^k - \bar{y}\|_G^2 \leq C\omega^{k/2},$$

where in the first inequality we have used $\tau < \frac{1}{L_f}$. To get (3.18), it suffices to ensure that

$\frac{C}{L_f}\omega^{k/2} \leq \epsilon$, which is equivalent to (since $\omega \in (0,1)$)

$$k \geq \frac{2\log(C/(L_f\epsilon))}{-\log(\omega)} = \frac{2\log(C/(L_f\epsilon))}{-\log\left((1+\delta)^{-1}\right)} = \frac{2\log(C/(L_f\epsilon))}{\log(1+\delta)} \geq \frac{2\log(C/(L_f\epsilon))}{\delta},$$

where we used $1 + \log\delta \leq \delta \quad \forall \delta > 0$. The results of the R-linear rate for dual sequence $\|y^k - \bar{y}\|_G$ follows similarly. This completes the proof. $\qquad \square$

The results we obtain in Theorem 3.1 and Corollary 3.2 show that the convergence rates depend on the parameter δ which also depends on the problem's data, the controlling parameters and a free parameter. It is desirable to minimize the theoretical bound $\frac{1}{1+\delta}$ of the Q-linear rate in (3.10) or alternatively to maximize δ. Here we consider the choice of the free parameter $\alpha > 1$ and the controlling parameters τ and σ which minimize the Q-linear rate upper bound given in (3.10).

One observes that δ cannot grow without bound by any choice of the controlling parameters and the free parameter, thus $\delta \leq \delta_m$ where δ_m is finite. First we set the other parameter $\sigma = \frac{1}{\tau\lambda_{max}(\mathcal{A}^T\mathcal{A})}$, and then we maximize δ with respect to the step size parameter τ.

For convenience, denote the condition number of $\mathcal{A}$ by

$$\kappa_{\mathcal{A}} := \sigma^2_{max}(\mathcal{A})/\sigma^2_{min}(\mathcal{A}) \equiv \lambda_{max}(\mathcal{A}^T\mathcal{A})/\lambda_{min}(\mathcal{A}^T\mathcal{A}).$$

We also denote the condition modulus of f by

$$\kappa_f \equiv L_f/\mu.$$

Set $\sigma = \frac{1}{\tau\lambda_{\max}(\mathcal{A}^T\mathcal{A})}$ in (3.11), this yields the expression

$$\delta = \min\left\{\frac{(\alpha-1)(1-\tau L_f)\kappa_{\mathcal{A}}^{-1}}{\alpha}, \frac{\mu\kappa_{\mathcal{A}}^{-1}}{\alpha\tau L_f^2 + \frac{\kappa_{\mathcal{A}}^{-1}}{\tau}}\right\}. \tag{3.20}$$

The parameter $\tau \in (0, \frac{1}{L_f})$ which can maximize δ in (3.20) can be found only from the second term inside the minimization above, and it evaluates to:

$$\tau_m = \frac{1}{\sqrt{\kappa_{\mathcal{A}}\alpha}L_f} =: \frac{1}{\rho L_f}, \tag{3.21}$$

19

where we define the quantity $\rho = \sqrt{\kappa_{\mathcal{A}}\alpha}$. Evaluating δ at τ_m, we get

$$\delta_m = \delta(\tau_m) = \min\left\{\frac{(\alpha - 1)(1 - \kappa_{\mathcal{A}}^{-1/2}\alpha^{-1/2})}{\alpha\kappa_{\mathcal{A}}}, \frac{1}{2\alpha^{1/2}\kappa_{\mathcal{A}}^{1/2}\kappa_f}\right\}$$
$$= \min\left\{\frac{(\rho^2 - \kappa_{\mathcal{A}})(1 - \rho^{-1})}{\rho^2\kappa_{\mathcal{A}}}, \frac{1}{2\rho\kappa_f}\right\}$$

The first term is monotonically increasing and the second terms is monotonically decreasing with respect to ρ, where $\rho > \sqrt{\kappa_{\mathcal{A}}}$. To maximize δ, we choose $\rho > \sqrt{\kappa_{\mathcal{A}}}$ (which corresponds to $\alpha > 1$) such that the two terms are equal. Some calculations show that this value can be found from the real solution of the cubic equation

$$\rho^3 - \left(1 + \frac{\kappa_{\mathcal{A}}}{2\kappa_f}\right)\rho^2 - \kappa_{\mathcal{A}}\rho + \kappa_{\mathcal{A}} = 0,$$

where $\rho > \sqrt{\kappa_{\mathcal{A}}}$. For specific values $\kappa_{\mathcal{A}}, \kappa_f \geq 1$ the root can be found by the bisection method over an appropriate interval, for example, when $\kappa_{\mathcal{A}} = 1, \kappa_f = 1$, the root is found in the interval $(1, 2]$ at $\rho \simeq 1.744$, then we may choose $\tau_m \simeq \frac{1}{1.744L_f} \simeq 0.57/L_f$. In addition when $\kappa_f \gg \kappa_{\mathcal{A}}$, we benefit from taking the gradient step size τ to be very close to $1/L_f$ as ρ tends to 1 in that case.

Overall, the value of δ_m decreases as $\kappa_{\mathcal{A}}$ or κ_f increase, which affects the practical performance of the algorithm. We thus arrive at the very natural conclusion that the smaller the condition numbers of $\kappa_{\mathcal{A}}$ and κ_f, the better the performance of the algorithm.

4 Computational Examples

We now illustrate the theoretical results with numerical experiments from image processing. We study two convex optimization problems in signal processing and imaging: image denoising with L_2-discrepancy and TV regularization, followed by a study of simultaneous image deconvolution and denoising with statistical multisresolution side constraints.

4.1 Total Variation - L₂-discrepancy and TV regularization

We test PAPC for a discrete L_2-TV signal denoising problem introduced in [18]. We apply the discretized variational problem of the total variation based signal denoising model according to

$$\min_{x \in X} \lambda \|\nabla x\|_1 + \frac{1}{2}\|x - b\|_2^2, \tag{4.1}$$

where X is a finite dimensional vector space equipped with the standard scalar product

$$\langle x, x' \rangle = \sum_i x_i x'_i, \qquad \forall x, x' \in X.$$

The gradient $\nabla : X \to Y$ is a vector in the vector space $Y = X$ and is defined by forward finite-difference operator with Dirichlet boundary conditions, $b \in X$ is the noisy signal, and $\lambda > 0$ is a positive parameter. The minimization problem (4.1) can be written as a saddle point problem

$$\min_{x \in X} \max_{y \in Y} \langle \mathcal{L}x, y \rangle + F(x) - G(y)$$

with primal variable $x \in X$, dual variable $y \in Y$, $F(x) = \frac{1}{2}\|x - b\|^2$, $G(y) = \delta_P(y)$, where P is given by the convex set $P = \{y : \|y\|_\infty \le \lambda\}$ and $\|y\|_\infty$ denotes the discrete maximum norm and the function δ_P denotes the indicator function of the set P, i.e.,

$$\delta_P(y) = \begin{cases} 0 & \text{if } y \in P, \\ \infty & \text{otherwise,} \end{cases}$$

and $\mathcal{L} = \nabla$. The adjoint ∇^* is the negative divergence, i.e., the unique linear mapping $\text{div} : Y \to X$ which satisfies $\langle x, \nabla y \rangle_Y = \langle x, \nabla^* y \rangle_X = -\langle x, \text{div } y \rangle_X$ for all $x \in X, y \in Y$. Finite differences are used for the discrete operator ∇ and its adjoint operator $\nabla^* = -\text{div}$ with the Dirichlet boundary conditions:

$$(\partial x)_i = \begin{cases} x_{i+1} - x_i & \text{if } 1 \le i < n \\ -x_n & \text{if } i = n \end{cases}, \qquad (\partial^* x)_i = \begin{cases} x_1 & \text{if } i = 1, \\ x_i - x_{i-1} & \text{if } 1 < i < n, \\ x_n - x_{n-1} & \text{if } i = n. \end{cases}$$

In each iteration of the PAPC algorithm we must evaluate the gradient of a smooth (strongly) convex function $F(\cdot)$ as well as the resolvent of G

$$(I + \sigma \partial G)^{-1}(y) = \operatorname*{argmin}_{y' \in Y} \left\{ \frac{1}{2}\|y' - y\|^2 + \mathcal{I}_{\|y\|_\infty \le \lambda}(y') \right\} = \frac{y}{\max\{1, |y|/\lambda\}}.$$

The bound on the square of norm of the operator ∇ (resp. div) in the one dimensional case,

$$\|\nabla\|^2 = \|\text{div}\|^2 \le \sum_{i=1}^{m-1}(x_i - x_{i+1})^2 \le 2\sum_{i=1}^{m-1}(x_i^2 + x_{i+1}^2) \le 4\sum_{i=1}^{m} x_i^2 \le 4.$$

21

Since the term $F(\cdot)$ is strongly convex and continuously differentiable with Lipschitz continuous gradient, and the operator in the bilinear term satisfies $\mathcal{L}^T \mathcal{L} = -\Delta$ (negative of the Laplacian operator) is positive definite under the Dirichlet boundary condition (see Appendix A), we obtain the expected R-linear rate for the sequence $\{x^k\}_{k \in \mathbb{N}}$ based on the theory developed in Section 3.

We illustrate the linear convergence of the method for finding the solution of the L$_2$-total-variation denoising problem in $\mathbb{R}^N$ with discretization $N = 256$. The noisy image is perturbed by an additive Gaussian noise with standard deviation 0.03. The regularization parameter λ was set to 0.05. The denoising result is shown in Figure 1(c). In the numerical experiments, we monitor the steps $\|u^{k+1} - u^k\|_H$, where $\{u\}_{k \in \mathbb{N}}$ is the primal-dual PAPC sequence. The a posteriori upper bound on the distance of the kth iterate to the true solution is $\|u^k - \hat{u}\|_H \leq \frac{c}{1-c}\|u^k - u^{k-1}\|_H$. With an estimated convergence rate of $c = 0.96$ for the run with $\tau = 0.05$ and $c = 0.985$ for the run with $\tau = 0.9$, this corresponds to an a posteriori upper estimate of the pointwise error at the iteration $k = 220$ of $1.0 * 10^{-6}$, or about five digits of accuracy at each pixel, for the run with $\tau = 0.05$, and $2.1 * 10^{-5}$, or about 4 digits of accuracy at each pixel, for the run with $\tau = 0.9$.
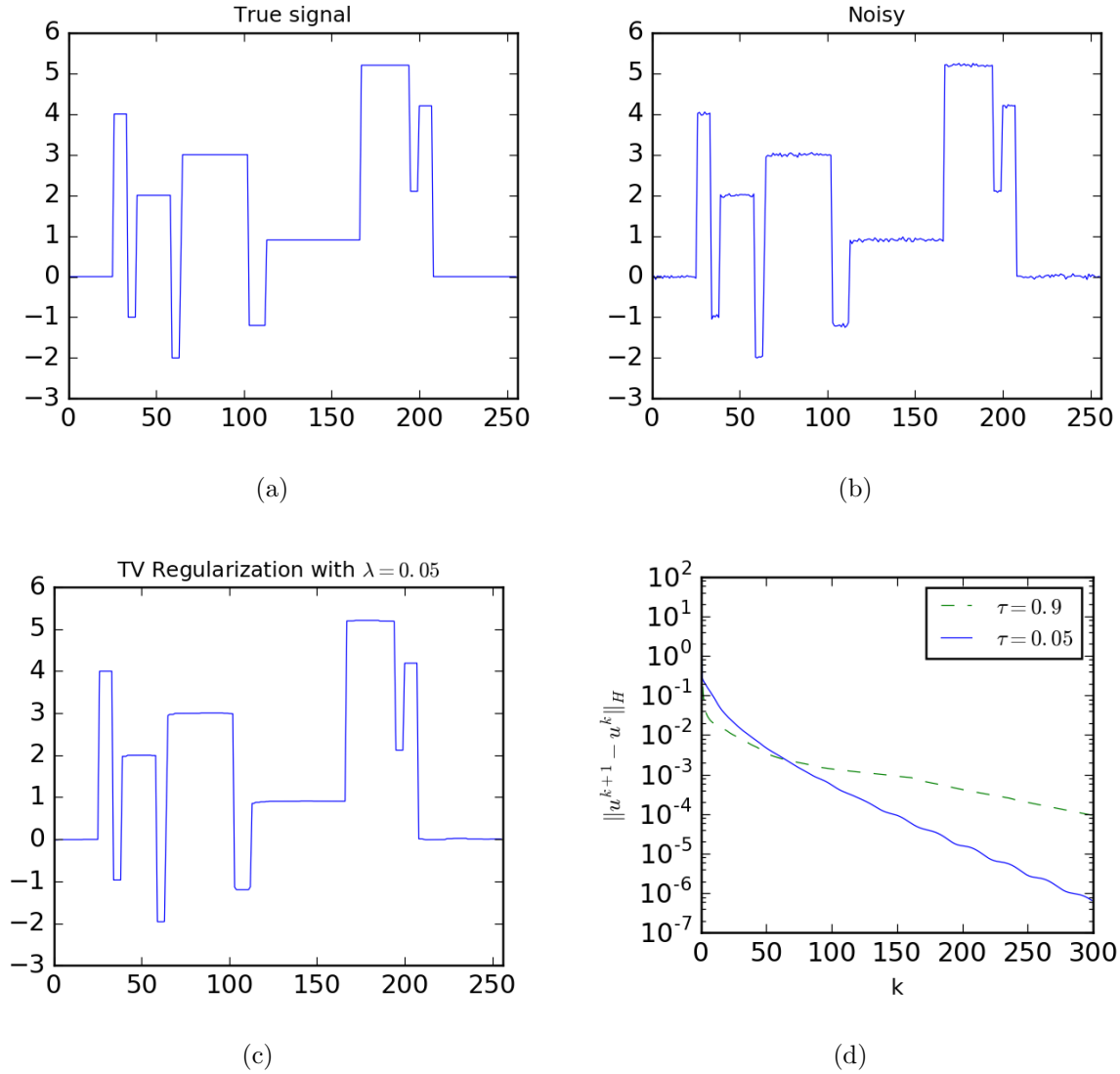
Figure 1: Performance evaluation of the PAPC algorithm for L_2-TV denoising. (a) Original signal. (b) Noisy signal with additive Gaussian noise, standard deviation 0.03 (c) The recovered signal using TV regularization with $\lambda = 0.05$. (d) The a posteriori error bound for linearly convergent iteration $\|u^{k+1} - u^k\|_H$ of the associated sequence as function of the number of iterations. The x-axis represents the number of iterations, and the y-axis represent the distance of consecutive iterates $\|u^{k+1} - u^k\|_H$ in logarithmic scale.

4.2 Statistical Multiresolution Analysis

In this subsection we apply the spatially-adaptive method for signal and image reconstruction that is based on Statistical MultiResolution Estimation (SMRE) as introduced in [2, 10, 9]. The variational manifestation of statistical multiresulution estimation is an optimization problem whose solution has a quantitative/statistical interpretation. We are not aware of any other variational image processing models whose solutions have such scientific content. The relevance to the theory presented in the previous sections is that, if one cannot estimate the distance of an iterate to a solution of the SMRE variational problem, then the statistical significance of the iterate generated by the algorithm cannot be determined – merely knowing that the sequence converges, or even knowing that the objective value converges at a particular rate, is useless information and empty of any scientific content.

In [1] a portion of the sequence (the dual sequence) of the alternating directions method of multipliers was shown to be locally linearly convergent, which yielded error estimates on solutions to iteratively regularized subproblems. We will give more specifics on the implementation below, but the advancement of the PAPC method in comparison to the ADMM method studied in [1] is two-fold: first, we can address a wider variety of regularizing objective functions and, secondly, the PAPC is dramatically more efficient, requiring less than $1/100$ of the time to achieve the kind of accuracy achieved in [1].

The variational problem involves the construction of an estimator for an unknown *true* signal by minimization of a convex functional J over a convex set that is determined by the statistical extreme value behavior of the residual. More precisely, we want to reconstruct the estimator $\bar{x}$ of the observed signal b that is computed as solutions of the convex optimization problem:

$$\inf_{x \in X} J(x) \quad \text{s.t} \quad \max_{s \in \mathcal{S}} \left| \sum_{\nu \in \mathcal{G}} \omega^s (Ax - b)_\nu \right| \leq q, \tag{4.2}$$

where $J : X \to \mathbb{R}$ denotes a regularization functional, which incorporates a priori knowledge on the unknown signal $\hat{x}$ such as smoothness, X is finite dimensional vector space, $A : X \to X$ is some linear mapping, $\mathcal{S}$ denotes a system of subsets of the grid $\mathcal{G}$ over X, and $\{w^s : s \in \mathcal{S}\}$ is a set of positive weights on the grid which are typically normalized indicator functions on $s \in \mathcal{S}$. The constant q serves as a regularization parameter which governs the trade-off between regularity and the fit to data of the reconstruction.

Solutions to problem 4.2 with SMRE constraints then have statistical content. We are then able to obtain quantitative (i.e., statistical) information about the iterates of the algorithm based on error bounds to solutions to 4.2 made possible by the convergence result

obtained in Theorem 3.1 and Corollary 3.2. The problem can be rewritten as a saddle point problem as follows:

$$\min_{x \in X} \max_{y \in Y} \{K(x, y) := J(x) + \langle x, \mathcal{A}y \rangle - G^*(y)\} \tag{$\mathcal{M}$}$$

where $y = (y_1, \ldots, y_{|\mathcal{S}|})$, $y_s \in X$, $s = 1, \ldots, |\mathcal{S}|$ $\mathcal{A}y = \sum_{s=1}^{|\mathcal{S}|} A^T y_i$ and G^* is the conjugate function of the sum of indicator functions $G(y) = \sum_s \delta_{C_s}(y)$ where each $C_s = \{y : |\sum_{\nu \in \mathcal{G}} \omega^s (y - b)| \leq q\}$. Using the Moreau's identity (1.3), the prox-mapping is evaluated in (2.9c) for each constraint by

$$y_s^k = \text{prox}_\sigma^{\delta_{C_s}^*} (y_s^{k-1} + \sigma A p^k) = y_s^{k-1} + \sigma A p^k - \sigma P_{C_s} \left(\frac{y_s^{k-1} + \sigma A p^k}{\sigma} \right), \qquad s = 1, \ldots, |\mathcal{S}|.$$

The proximal parameter is a function of τ and given by $\sigma = 1/(\tau \|A^T A\|_2)$. More details in [8, Sect. 4.1].

Our approach here is to use $J(x) = \frac{1}{2}\|\nabla x\|^2$ and the smoothed TV norm. Because the gradient ∇ has a nontrivial kernel, we cannot guarantee generically that the objective J is pointwise quadratically supportable. We recover strong coercivity by imposing appropriate boundary conditions. In the two-dimensional example, this is combined with a Huber function to demonstrate the full extent of the theory.

4.2.1 One-dimensional SMRE – synthetic data

In this example, we consider a one dimensional signal as shown in Figure 3(a), and the corresponding noisy data Figure 3(b) with $n = 512$ data points. The noisy data was generated by adding independently and identically distributed Gaussian random noise with standard deviation 0.02 to each measured data point. In this problem we consider only denoising, that is, the operator A in (4.2) is the identity operator. The qualitative objective is $J(x) = \frac{1}{2}\|\nabla x\|^2$.

The constraints are applied on all windows consisting of all intervals of length between 1 and L pixels. The weights $\{w_j \in \mathbb{R}^n\}$ are scaled so that the windows are normalized, and $\mathcal{S}$ is the index set corresponding to all collections of successive pixels in $\{1, 2, ..., n\}$ of length from 1 to l, $1 \leq l \leq L$, where L represents the total number of multiresolution levels used in the construction. For a signal length $n = 512$ and $L = 10$ multiresultion levels, the number constraints are $|\mathcal{S}| = 5075$. The constraints/windows have a separable block structure on nonoverlapping windows which can be exploited for efficient implementation. This allows us to perform the parallel block updates in the dual space using $\frac{1}{2}L(L + 1)$ dual variables
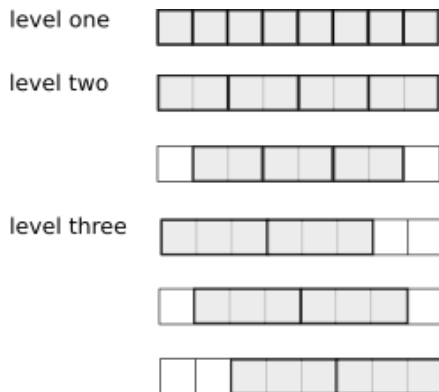
Figure 2: Schematic window tiling of the multiscale constraints in the 1D case.

of length n. Figure 2 shows the schematic tiling of the nonoverlapping window constraints using lengths varying up to $L = 3$.

The objective function, $J(x) = \frac{1}{2}\|\nabla x\|^2$, is strongly convex if for every bounded neighborhood of a solution x, q is small enough such that there are no constant functions over the support of 3 pixels around x_i which satisfy the constraint $|x_i - b_i| \leq q$, $1 \leq i \leq n$. We set the error bound q starting with q_0 corresponding to 3 standard deviations for the first level and for each subsequent multiresolution it is decreased based on the window length by a factor $q_j = q_0 f^{l-1}$ is where $f \in (0, 1)$ is a scaling factor and $l \in \{1, \ldots, L\}$ is the scaling level. The last computed iterate is shown in Figure 3(c) using one and ten multiresolution levels demonstrating the advantage of the SMRE approach. With an estimated convergence rate of $c = 0.999$ for the run with $\tau = 0.2$ and $c = 0.9985$ for the run with $\tau = 0.02$, this corresponds to an a posteriori upper estimate of the error at iteration $k = 1200$ of $1.83*10^{-3}$, or two digits of accuracy at each pixel, and $1.09 * 10^{-4}$, or three digits of accuracy per pixel, respectively.
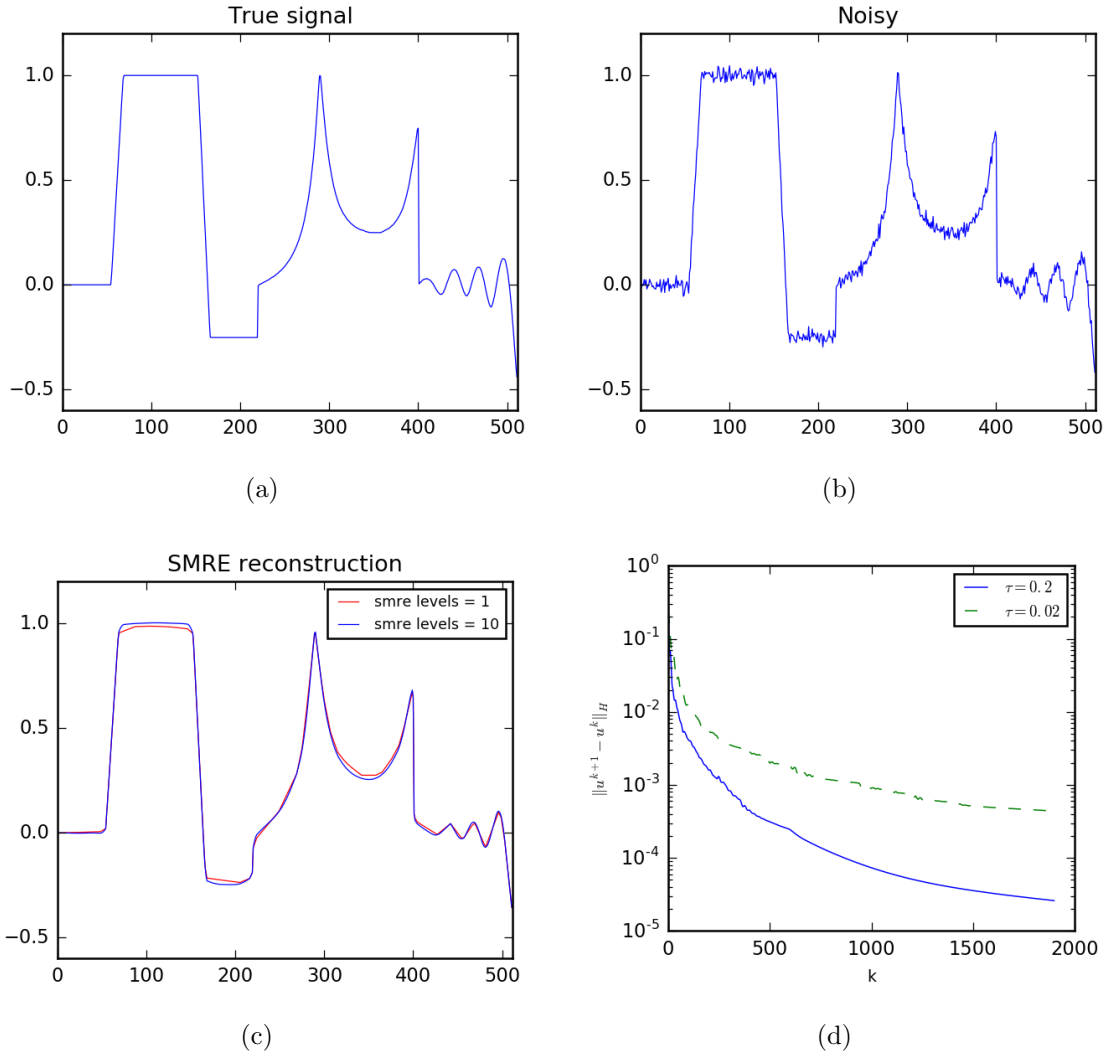
Figure 3: Results of statistical multiresolution reconstruction of one dimensional signal with $L = 1$ and $L = 10$ levels using q corresponding to 3 standard deviations and scaling factor $f = 0.93$. (a) The original signal. (b) Noisy with additive Gaussian noise, standard deviation 0.02 (c) The multiresolution signal reconstruction with number of levels $L = 1$ and $L = 10$. (d) The a posteriori error bound for linearly convergent iterations given by the distance of consecutive iterates $\|u^{k+1} - u^k\|_H$ of the associated sequence as function of the no. of iterations. The x-axis represents the number of iterations, and the y-axis represent the distance of consecutive iterates $\|u^{k+1} - u^k\|_H$ in logarithmic scale.

27

4.2.2   Two-dimensional SMRE – Image Laboratory Data

We next consider the performance of our method in image reconstruction problems for Stimulated Emission Depletion (STED) microscopy experiment conducted at the Laser-Laboratorium Göttingen examining tubulin. The SMRE model problem (4.2) described at length for the 1D domain can be easily extended for an image deconvolution and denoising in the 2D case, where the linear operator A is the convolution using a point spread function.

Here, we also consider the smooth approximation of total variation (TV) functional as the qualitative objective. The TV functional is defined by $\|\nabla x\|_1$ where $x \in X = \mathbb{R}^{n \times n}$ and is often used for image filtering and restoration. However, this function is non-smooth where $|\nabla x| = 0$, thus in order to make the derivative-based methods possible we consider a smoothed approximation of the TV functional known as the so-called Huber approximation.

The Huber loss function is defined as follows:

$$\|x\|_{1,\alpha} = \sum_{i,j} \phi_\alpha(x_{i,j}), \qquad \phi_\alpha(t) = \begin{cases} \frac{t^2}{2\alpha} & \text{if } |t| \leq \alpha \\ |t| - \frac{\alpha}{2} & \text{if } |t| > \alpha, \end{cases} \tag{4.3}$$

where $\alpha > 0$ is a small parameter defining the trade-off between quadratic regularization (for small values) and total variation regularization (for larger values). The function ϕ is smooth with $\frac{1}{\alpha}$-Lipschitz continuous derivative and its derivative is given by

$$\phi'_\alpha(t) = \begin{cases} \frac{t}{\alpha} & \text{if } |t| \leq \alpha \\ \text{sgn}(t) & \text{if } |t| > \alpha. \end{cases} \tag{4.4}$$

The objective functional that is used in (4.2) is $J(x) = \|\nabla x\|_{1,\alpha}$, where $x \in X = \mathbb{R}^{n \times n}$, and ∇ is the vector in the vector space $Y = X \times X$. For discretization of $\nabla : X \to Y$, we use standard finite differences with Neumann boundary conditions,

$$(\nabla x)_{i,j} = \begin{pmatrix} (\nabla_1 x)_{i,j} \\ (\nabla_2 x)_{i,j} \end{pmatrix}, \qquad 1 \leq i, j \leq n,$$

where

$$(\nabla_1 x) = \begin{cases} x_{i+1,j} - x_{i,j} & \text{if } 1 \leq i < n-1 \\ 0 & \text{if } i = n-1 \end{cases} \qquad (\nabla_2 x) = \begin{cases} x_{i,j+1} - x_{i,j} & \text{if } 1 \leq j < n-1 \\ 0 & \text{if } j = n-1. \end{cases}$$

The Laplacian computed at the pixel (i, j) is given using the standard finite differences above as $(\Delta x)_{i,j} = 4x_{i,j} - x_{i-1,j} - x_{i+1,j} - x_{i,j+1} - x_{i,j-1}$. The objective $J(x)$ is continuously differentiable with Lipschitz gradient constant $L_J = \|\nabla\|^2/\alpha \leq 8/\alpha$. This function is pointwise quadratically supportable at solutions as long as on the small bounded neighborhood of a solution $\bar{x}$, where $|(\Delta \bar{x})_{i,j}| < \alpha$, q is small enough that there are no constant functions over the support of 3×3 around $\bar{x}_{i,j}$ which satisfy the constraints $|(A\bar{x})_{i,j} - b_{i,j}| \leq q$, $1 \leq i, j \leq n$. This much we must assume.

We demonstrate our results of reconstruction a close-up image shown in Figure 4 of dimension size $n \times n$ where $n = 64$ using the statistical multiresolution levels approach. The confidence level q was set to 0.07. The graphs in Figure 5 show the a posteriori error bound results of the algorithm's performance with quadratic objective $\frac{1}{2}\|\nabla x\|^2$ and the Huber smooth approximation $\|\nabla x\|_{1,\alpha}$, where $\alpha = 0.25$. The step size of the gradient step of both runs was set to $\tau = 0.02$. The quadratic model achieves a better rate of convergence. With an estimated convergence rate of $c = 0.9993$ for the Huber objective this corresponds to an a posteriori upper estimate of the error at iteration $k = 800$ of $2.4 * 10^{-3}$. With an estimated convergence rate of $c = 0.9962$ for the quadratic objective function this corresponds to an a posteriori upper estimate of the error at iteration $k = 800$ of $1.5 * 10^{-3}$ – about two digits of accuracy at each pixel.
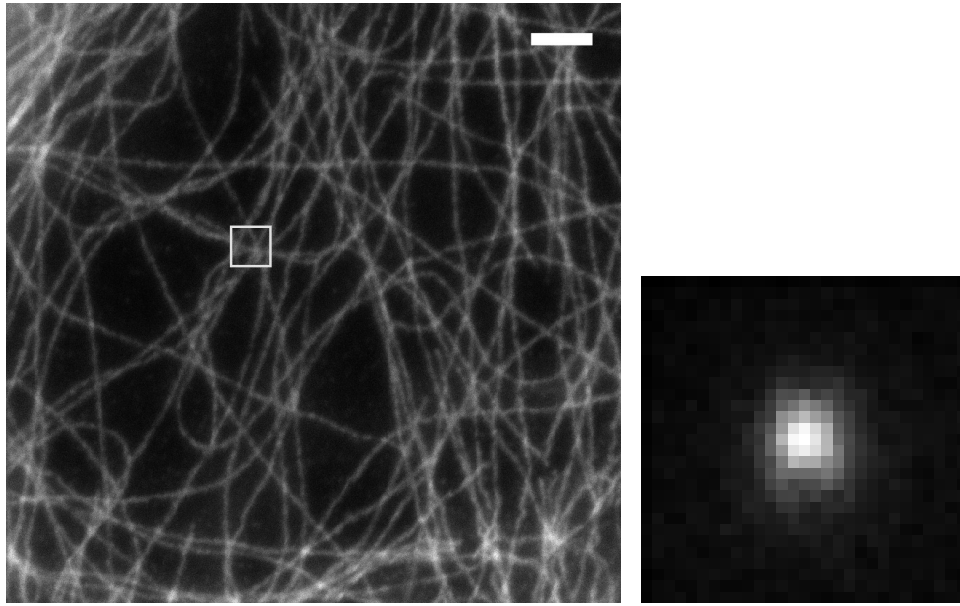
(a)   (b)
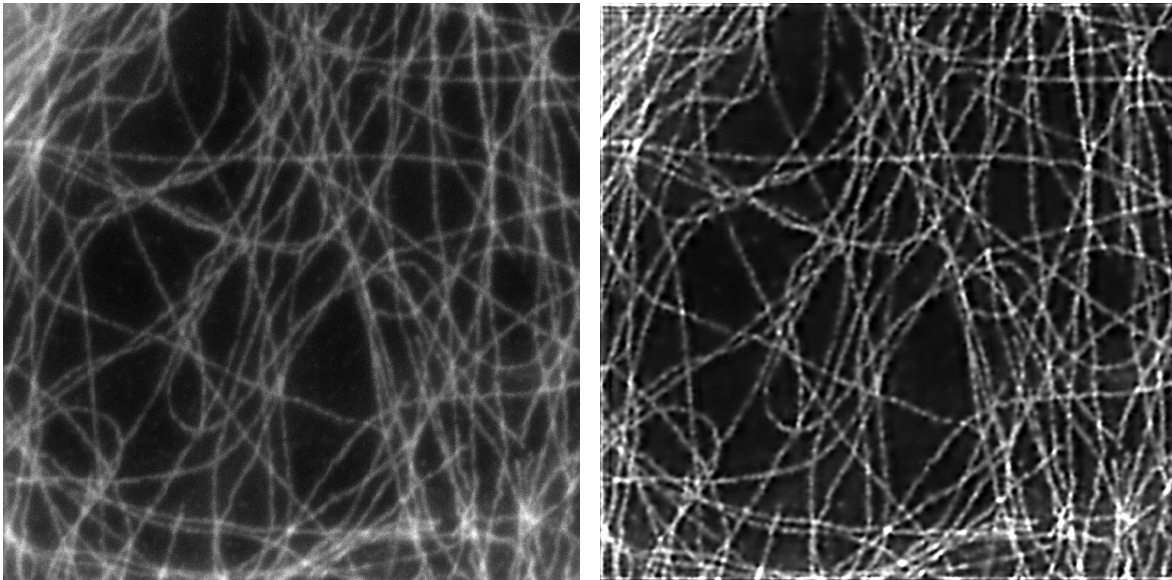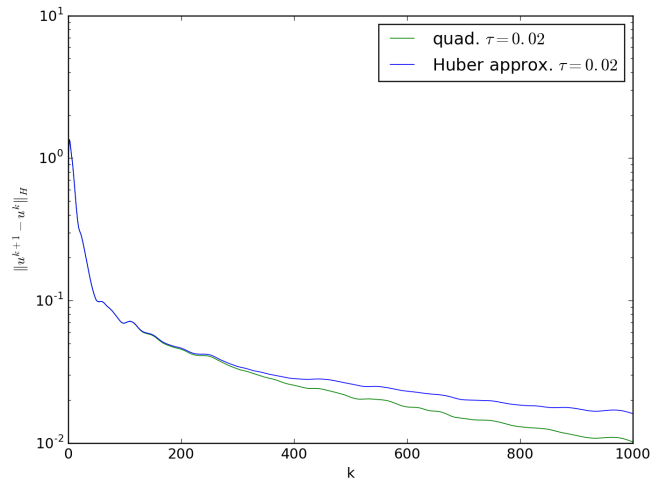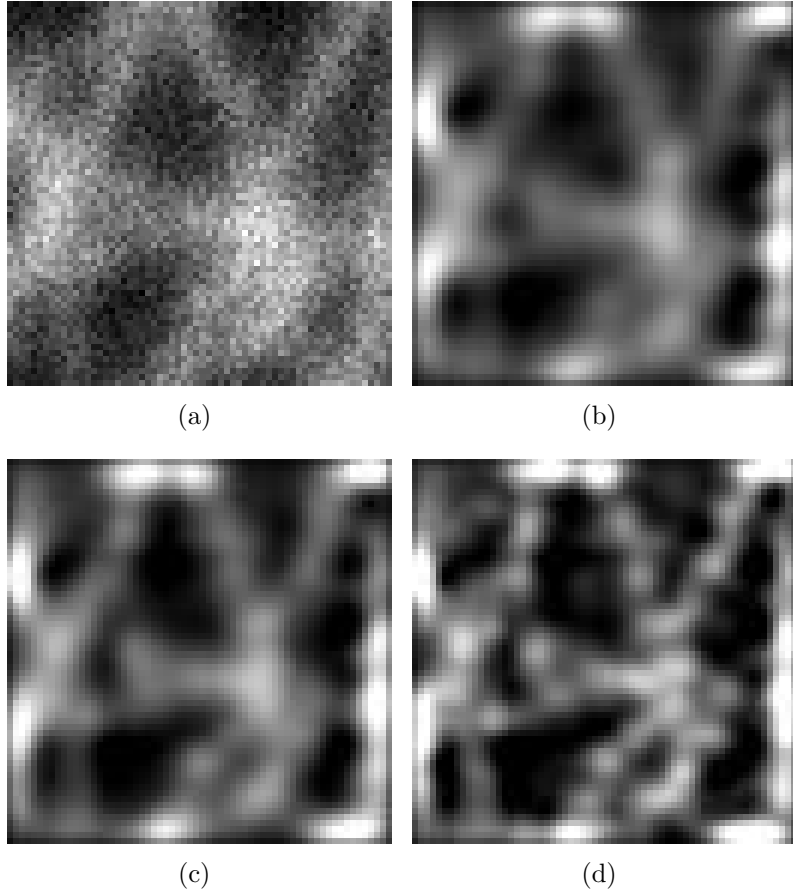

(c)   (d)

Figure 4: (a) Original data (STED image of Tubulin) with a close-up box to be processed, (b) PSF image, (c) full image deconvolution and denoising using the statistical multiresolution method with $L = 3$ with the objective functional

(a)   (b)

(c)   (d)

(e)

Figure 5: Results of image denoising and deconvolution using statistical multiresolution estimation with three levels. (a)-(d) depicts the evolution of the image during reconstruction after 0, 50, 150, 1000 iterations with $\|x\|_{0.25,1}$. (e) The primal sequence increment $\|u^{k+1} - u^k\|_H$ of the associated sequence as function of the number of iterations.

A   Appendix

It is well known that the minimum eigenvalue of the negative of the Laplacian is positive under Dirichlet boundary condition. Consider the eigenvalues problem of the negative of the Laplacian operator in 1D

$$\begin{cases} -\Delta x = \lambda x, \\ x_0 = x_{n+1} = 0, \quad x \in \mathbb{R}^n \end{cases} \tag{A.1}$$

The Laplacian operator (here the second order derivative) can be discretized using finite differencing method, and here we consider using the standard first order discretizations $(-\Delta x)_j = (-x_{j-1} + 2x_j - x_{j+1})$, $j = 1, \ldots, n$. The last expression needs to be adjusted at the boundary points based on (A.1).

Let us use the discretized solution based on the continuous case eigenvector $x_{kj} := (x_k)_j = \sin\left(\frac{k\pi j}{n+1}\right)$ which satisfies $(x_k)_0 = (x_k)_{n+1} = 0$, and we show that it is indeed the eigenvector k computed for the component j (without the normalization constant). We plug in to the Laplacian matrix to get:

$$\begin{aligned} (-\Delta x_k)_j &= -\sin\left(\frac{k\pi(j-1)}{n+1}\right) + 2\sin\left(\frac{k\pi j}{n+1}\right) - \sin\left(\frac{k\pi(j-1)}{n+1}\right) \\ &= \left[2 - 2\cos\left(\frac{k\pi}{n+1}\right)\right]\sin\left(\frac{k\pi j}{n+1}\right) \\ &= 4\sin^2\left(\frac{k\pi}{2n+2}\right)\sin\left(\frac{k\pi j}{n+1}\right). \end{aligned}$$

This shows that x_{kj} is eigenvector with eigenvalue $4\sin^2\left(\frac{k\pi}{2(n+1)}\right)$. The minimum eigenvalue of the nagative of the Laplacian under Dirichlet boundary condition is thus

$$\min_{\substack{x \neq 0 \\ x_0 = 0, x_{n+1} = 0}} \frac{-x^T \Delta x}{x^T x} = \frac{\sum_{j=1}^n 4\sin^2\left(\frac{\pi}{2n+2}\right)\sin\left(\frac{\pi j}{n+1}\right)^2}{\sum_{j=1}^n \sin\left(\frac{\pi j}{n+1}\right)^2} = 4\sin^2\left(\frac{\pi}{2n+2}\right).$$

Acknowledgments

References

[1] T. Aspelmeier, C Charitha, and D Russell Luke. Local linear convergence of the admm/douglas–rachford algorithms without strong convexity and application to statistical imaging. *SIAM J. Imaging Science, to appear*, 2016.

[2] T. Aspelmeier, A. Egner, and A. Munk. Modern statistical challenges in high-resolution fluorescence microscopy. *Annual Review of Statistics and Its Application*, 2:163–202, 2015.

[3] A. Auslender and M. Teboulle. *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. Springer Monographs in Mathematics. New York: Springer, 2003.

[4] J.M. Borwein and J. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*, volume 109 of *Encyclopedias in Mathematics*. Cambridge University Press, New York, 2010.

[5] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, pages 1–26, 2010.

[6] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016.

[7] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

[8] Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.

[9] K. Frick, P. Marnitz, and A. Munk. Statistical multiresolution estimation for variational imaging: With an application in poisson-biophotonics. *Journal of Mathematical Imaging and Vision*, 46(3):370–387, 2013.

[10] K. Frick, P. Marnitz and A. Munk Statistical multiresolution dantzig estimation in imaging: fundamental concepts and algorithmic framework. *Electronic Journal of Statistics*, 6:231–268, 2012.

[11] S. W. Hell and J. Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics letters*, 19(11):780–782, 1994.

[12] T. A. Klar, S. Jakobs, M. Dyba, A. Egner, and S. W. Hell. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proceedings of the National Academy of Sciences*, 97(15):8206–8210, 2000.

[13] D. R. Luke, Nguyen H. Thao, and M. T. Tam. Quantitative convergence analysis of iterated expansive, set-valued mappings. http://arxiv.org/abs/1605.05725, May 2016.

[14] J.J. Moreau. Proximité et dualité dans un espace Hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

[15] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[16] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[17] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundl. Math. Wiss.* Springer-Verlag, Berlin, 1998.

[18] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[19] O. Scherzer, editor. *Handbook of Mathematical Methods in Imaging.* Springer-Verlag, 2 edition, 2015.

[20] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM Journal on Optimization*, 24(1):269–297, 2014.

[21] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning.* Mit Press, 2011.