Cell Detection by Functional Inverse Diffusion and Non-negative Group Sparsity—Part II: Proximal Optimization and Performance Evaluation

Pol del Aguila Pla, *Student Member, IEEE*, and Joakim Jaldén, *Senior Member, IEEE*

***Abstract*—In this two-part paper, we present a novel framework and methodology to analyze data from certain image-based biochemical assays, e.g., ELISPOT and Fluorospot assays. In this second part, we focus on our algorithmic contributions. We provide an algorithm for functional inverse diffusion that solves the variational problem we posed in Part I. As part of the derivation of this algorithm, we present the proximal operator for the non-negative group-sparsity regularizer, which is a novel result that is of interest in itself, also in comparison to previous results on the proximal operator of a sum of functions. We then present a discretized approximated implementation of our algorithm and evaluate it both in terms of operational cell-detection metrics and in terms of distributional optimal-transport metrics.**

***Index Terms*—Proximal operator, non-negative group sparsity, functional optimization, biomedical imaging, source localization.**

I. Introduction

Source localization (SL) arises in application fields in which a number of point-sources emit some measurable signal, e.g., chemical compounds [1]–[12], sound [13], [14], light [15], [16] or heat [17], [18], and one wants to recover their location. Typically, the measured signal is a map of these locations observed through a linear operator, such as convolution [10], [13] or diffusion [2], [3], [9]. This is the second part of a paper that considers SL in the context of cell detection in image-based biochemical assays. In this case, the source locations are explicit in the source density rate, the reactive term in a reaction-diffusion-adsorption-desorption system, from which a single image of the adsorped density at the end of the experiment is measured. For more details on the exact setup, its biological application or the physics involved, see Part I [19, Section II].

In Part I [19, Section III], we proposed the following optimization problem to detect particle-generating (active) cells in this setting,

$$\min_{a\in\mathcal{A}}\left[\left\|Aa-d_{\mathrm{obs}}\right\|_{\mathcal{D}}^{2}+\underbrace{\delta_{\mathcal{A}_{+}}(a)+\lambda\overbrace{\left\|\left\|\xi a_{\mathbf{r}}\right\|_{\mathrm{L}^{2}(\mathbb{R}_{+})}\right\|_{\mathrm{L}^{1}(\mathbb{R}^{2})}}^{f_{1}(a)}}_{f(a)}\right]. \quad (1)$$

Here, the non-negative quantity a we aim to recover is the post adsorption-desorption source density rate (PSDR). The PSDR is an equivalent to the source density rate (SDR) where the information on adsorption and desorption have been summarized. In fact, it characterizes the generation of particles across a plane, that we represent by the locations $\mathbf{r}\in\mathbb{R}^2$, and a third non-negative dimension $\sigma\geq 0$ that expresses the distance each generated particle has diffused from its origin. Here, non-negative group-sparsity, induced by the regularizer f in (1), plays a fundamental role. This is because a certain form of grouping [20] is key for the end application, i.e., cell detection, but a has to remain non-negative at all times to preserve its physical meaning. In particular, it is important that the different values of $a_{\mathbf{r}}(\sigma)=a(\mathbf{r},\sigma)$ across the different distances σ for a certain position $\mathbf{r}$ are grouped, because they represent the same potential active cell generating particles which are captured either closer to the cell (low σ s) or further away (large σ s). To our knowledge, previous techniques for promoting group-sparsity, e.g., [21], [22], can not handle non-negativity constraints.

The Hilbert space where this PSDR lies is defined as $\mathcal{A}=\left\{a\in\mathrm{L}^2(\Omega):\operatorname{supp}(a)\subseteq\operatorname{supp}(\mu)\times[0,\sigma_{\max}]\right\}$, where $\Omega=\mathbb{R}^2\times\mathbb{R}_+$, μ is a $(0,1)$-indicator function of a bounded set $\operatorname{supp}(\mu)$ in which cells can physically lie, and $\sigma_{\max}=\sqrt{2DT}$ is given by the physical parameters of the assay. The image observation d_{obs} lies in a weighted $\mathrm{L}^2\left(\mathbb{R}^2\right)$ space defined as $\mathcal{D}=\left\{d:\mathbb{R}^2\to\mathbb{R}:(d|d)_{\mathcal{D}}<+\infty\right\}$, where $(d_1|d_2)_{\mathcal{D}}=\left(w^2\cdot|\cdot\right)_{\mathrm{L}^2(\mathbb{R}^2)}$ and $w\in\mathrm{L}_+^\infty\left(\mathbb{R}^2\right)$ is a non-negative bounded weighting function that penalizes errors at different locations according to sensor properties. The bounded linear operator $A\in\mathcal{L}(\mathcal{A},\mathcal{D})$ represents the forward diffusion process, that maps an a onto d_{obs}, and was derived in [19, Section II-B] as the mapping $a\mapsto\int_0^{\sigma_{\max}}G_\sigma a_\sigma\,\mathrm{d}\sigma$, where G_σ is the

Manuscript received September 21, 2017; revised March 31, 2018 and July 9, 2018; accepted August 24, 2018. Date of publication September 3, 2018; date of current version September 14, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark A. Davenport. This work was supported in part by Mabtech AB and in part by Grant 2015-04026 of the Swedish Research Council (VR). *(Corresponding author: Pol del Aguila Pla.)*

The Authors are with the Department of Information Science and Engineering, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm 11428, Sweden (e-mail: poldap@kth.se; jalden@kth.se).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material includes detailed derivations of some key steps and further experimental results. This material is 533 kB in size.

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.2018.2868256

convolutional operator with 2D rotationally invariant Gaussian kernel of standard deviation σ. The diffusion operator A was extensively characterized in Part I [19, Section III-B]. Finally, the group-sparsity regularizer includes a non-negative bounded weighting function $\xi \in \mathrm{L}^{\infty}_{+}\left[0, \sigma_{\max}\right]$ that allows incorporating further prior knowledge in terms of the relative importance of each value of σ.

In this paper, we derive an accelerated proximal gradient (APG) algorithm to solve (1). Namely, we combine the characterization of the diffusion operator A we presented in Part I [19, Section III-B] with the derivation of the proximal operator of the non-negative group-sparsity regularizer, i.e., of f in (1). Furthermore, we present an efficient implementation of a discretization of the resulting algorithm and provide thorough performance evaluation on synthetic data, complementing the real data example in Part I [19, Section V-A].

A. Proximal Operator of a Sum of Functions

Proximal methods for convex optimization [23]–[25] are now prevalent in the signal processing, inverse problems and machine learning communities [26]–[28]. This is mainly due to their first-order nature, i.e., that the intermediate variables they entail have at most the same dimensionality as the variable one seeks, and to their ability to handle complex, non-smooth shapes of the functional to optimize. Consequently, applications are characterized by high-dimensional parameters with rich structure and non-smooth penalizations.

In the most generic setting, the problems solved by these methods are of the form

$$\min_{x \in \mathcal{X}} \left[g(Bx) + f(x) \right] , \tag{2}$$

where $f : \mathcal{X} \to \bar{\mathbb{R}}$ and $g : \mathcal{G} \to \bar{\mathbb{R}}$ are proper, convex, and lower semi-continuous and the domains $\mathcal{X}$ and $\mathcal{G}$ are two real Hilbert spaces. On one hand, a smoothness assumption is made on g, namely, that it is Frchet differentiable in $\mathcal{G}$ and has a β^{-1}-Lipschitz continuous Frchet derivative $\nabla g : \mathcal{G} \to \mathcal{G}^{*}$ for some $\beta > 0$. On the other hand, no further structure is imposed on f, which can be non-smooth and discontinuous. Finally, the bounded linear operator $B \in \mathcal{L}\left(\mathcal{X}, \mathcal{G}\right)$ has an adjoint $B^{*} \in \mathcal{L}\left(\mathcal{G}, \mathcal{X}\right)$ and operator norm $\|B\|_{\mathcal{L}(\mathcal{X},\mathcal{G})}$.

The term *proximal* that encompasses these methods relates to the proximal operator of the function f, which is necessary for a fundamental step in the iterations defined by these algorithms. The proximal operator is a mapping $\mathrm{prox}_{\gamma f} : \mathcal{X} \to \mathcal{X}$ such that

$$\mathrm{prox}_{\gamma f}(x) = \arg\min_{y \in \mathcal{X}} \left[\|y - x\|^{2}_{\mathcal{X}} + 2\gamma f(y) \right] . \tag{3}$$

A case that generates special interest is that in which f is constructed as a sum of two non-smooth components [26], [28]–[31]. In particular, [29, Proposition 12] proved that if $\mathcal{X} = \mathbb{R}$ and $f = f_1 + \delta_{\mathcal{Z}}$, with $\delta_{\mathcal{Z}}$ the $(\infty, 0)$-indicator function of a closed convex subset $\mathcal{Z} \subset \mathcal{X}$, then $\mathrm{prox}_{f} = \mathrm{P}_{\mathcal{Z}} \circ \mathrm{prox}_{f_1}$, where $\circ$ represents composition and $\mathrm{P}_{\mathcal{Z}}$ the projection onto $\mathcal{Z}$. In the context of the derivation of the proximal operator of the non-negative group-sparsity regularizer f in (1), we provide a contrasting result. In particular, in the appendix to this paper, we prove that if $\mathcal{X} = \mathrm{L}^{2}$, $\mathcal{Z} = \mathcal{X}_{+}$, and $f_1 = \|\cdot\|_{\mathcal{X}}$, the inverse order applies, i.e., $\mathrm{prox}_{f} = \mathrm{prox}_{f_1} \circ \mathrm{P}_{\mathcal{Z}}$.[1] Combining this result with the separable sum property allows us to prove that this same order is applicable when $f = \lambda f_1 + \delta_{\mathcal{Z}}$ for some $\lambda \geq 0$, f_1 is the group-sparsity regularizer with non-overlapping groups, and $\mathcal{X}$ and $\mathcal{Z}$ are as above. Besides allowing us to solve (1), the proximal operator for the non-negative group sparsity regularizer facilitates the use of group-sparsity in other fields that inherently require non-negativity constraints, e.g., classification, text mining, environmetrics, speech recognition and computer vision [35], [36].

B. Notation

When sets and spaces of numbers are involved, we use either standard notation such as $\mathbb{R}_{+} = [0, +\infty)$, $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and $\bar{\mathbb{R}}_{+} = [0, +\infty]$ or capital non-Latin letters. When discussing locations in $\mathbb{R}^{2}$, we note them as bold face letters, e.g., $\mathbf{r} \in \mathbb{R}^{2}$.

When discussing functional sets and spaces, we use capital calligraphic notation, such as $\mathcal{X}$ for a generic normed space, $\|\cdot\|_{\mathcal{X}}$ for its norm, and $(\cdot|\cdot)_{\mathcal{X}}$ for its inner product if $\mathcal{X}$ is also a Hilbert space. For any functional space $\mathcal{X}$, $\mathcal{X}_{+} \subset \mathcal{X}$ is the cone of non-negative functionals, and for any functional f, f_{+} is its positive part, i.e., $f_{+}(y) = \max\{f(y), 0\}$, for any y in its domain $\mathcal{Y}$. The support of the functional f is written as $\mathrm{supp}(f) = \{y \in \mathcal{Y} : f(y) \neq 0\} \subset \mathcal{Y}$. For any set $\mathcal{Z} \subseteq \mathcal{X}$, its $(\infty, 0)$-indicator function is the function $\delta_{\mathcal{Z}} : \mathcal{X} \to \{0, +\infty\}$ such that $\delta_{\mathcal{Z}}(x) = 0$ if $x \in \mathcal{Z}$ and $\delta_{\mathcal{Z}}(x) = +\infty$ if $x \in \mathcal{Z}^{\mathsf{c}} = \mathcal{X} \setminus \mathcal{Z}$.

When discussing operators, if $\mathcal{Z}$ is some normed space, we write $\mathcal{L}(\mathcal{X}, \mathcal{Z})$ for the space of linear continuous operators from $\mathcal{X}$ to $\mathcal{Z}$, with norm $\|\cdot\|_{\mathcal{L}(\mathcal{X},\mathcal{Z})}$. We will note operators as A or B, e.g., $B \in \mathcal{L}\left(\mathcal{X}, \mathcal{Z}\right)$. Adjoints are noted as $B^{*} \in \mathcal{L}\left(\mathcal{Z}, \mathcal{X}\right)$.

When discussing matrices and tensors, the space of real M-by-N matrices for some $M, N \in \mathbb{N}$ is $\mathbb{T}\left(M, N\right)$, while its element-wise positive cone is $\mathbb{T}_{+}\left(M, N\right)$. For a specific matrix $\tilde{f} \in \mathbb{T}\left(M, N\right)$, we specify it as a group of its elements $\left(\tilde{f}_{m,n}\right)$ for $m \in \{1, 2, \ldots, M\}$ and $n \in \{1, 2, \ldots, N\}$. For tensors, we work analogously by adding appropriate indexes, e.g., $\tilde{f} \in \mathbb{T}\left(M, N, K\right)$ and $\left(\tilde{f}_{m,n,k}\right)$ for $k \in \{1, 2, \ldots, K\}$.

When presenting our statements, we refer to them as properties if they are not novel, but are necessary for clear exposition, lemmas if they contain minor novel contributions and theorems if they constitute major novel contributions.

II. Accelerated Proximal Gradient for Weighted Group-Sparse Regularized Inverse Diffusion

In this section, we propose to use the accelerated proximal gradient algorithm (APG algorithm or FISTA [23]) to solve (1). Because the APG algorithm is posed in generic Hilbert spaces, we do not need to discretize the problem in order to derive and pose our algorithm. In other words, the proposed

[1] After the acceptance of this paper, we were informed that this claim, made in our Lemma 4, was covered by the broader result [32, Proposition 2.2]. During the revision of this paper, [33], [34] also presented statements equivalent or encompassing Lemma 4.

Algorithm 1: Accelerated Proximal Gradient to find x_{opt} that solves (2) with function-value convergence rate $\mathcal{O}\left(i^{-2}\right)$. To simplify exposition, we identify $\mathcal{G}$ with its dual $\mathcal{G}^*$ and write $\nabla g(By)$ to refer to its representation in $\mathcal{G}$. We also note $\|B\|^2 = \|B\|^2_{\mathcal{L}(\mathcal{X},\mathcal{G})}$. Moreover, the ratio $\|B\|^2/\beta$ here is representing the best Lipschitz continuity constant for $\nabla\left(g \circ B\right)$, but can be replaced by any constant upper bound of this quantity and the convergence rate $\mathcal{O}\left(i^{-2}\right)$ will still be preserved.

```
Require: An initial $x^{(0)} \in \mathcal{X}$
Ensure: A solution $x_{\text{opt}} \in \mathcal{X}$ that solves (2)
$y^{(0)} \leftarrow x^{(0)}, i \leftarrow 0$
repeat
    $i \leftarrow i+1, \alpha \leftarrow \frac{t(i-1)-1}{t(i)}$
    $x^{(i)} \leftarrow \text{prox}_{\frac{\beta}{\|B\|^2} f}\left[y^{(i-1)} - \frac{\beta}{\|B\|^2} B^* \nabla g\left(By^{(i-1)}\right)\right]$
    $y^{(i)} \leftarrow x^{(i)} + \alpha\left(x^{(i)} - x^{(i-1)}\right)$
until convergence
$x_{\text{opt}} \leftarrow x^{(i)}$
```

algorithm will work directly on the abstract parameter space $\mathcal{A}$. For an introduction on optimization in function spaces, see [37]. Any implementation, however, will require some form of discretization. In our case, we choose the simple discretization presented in Part I [19, Section IV].

A. Accelerated Proximal Gradient

The APG algorithm, i.e., Algorithm 1, was proposed in [23] to solve (2) with $t : \mathbb{N} \rightarrow \mathbb{R}_+$ such that $t(i) = 1/2 + [1/4 + t^2(i-1)]^{1/2}$, $\forall i \geq 1$ and $t(0) = 1$. [23] proved that this algorithm yields a sequence of objective values $f(x_i) + g(Bx_i)$ with a convergence rate towards the minimum value of $\mathcal{O}\left(i^{-2}\right)$. [38] proposed modifying the update rule of the Nesterov acceleration term to $t(i) = (i + a - 1)/a, \forall i \geq 1$, for some $a > 2$. This modification preserves the convergence rate of the objective values, and additionally grants weak convergence of the iterates, i.e., $x_i \rightarrow x_{\text{opt}}$ weakly (see [37] for more on weak convergence). As discussed in [38], convergence is observed empirically with the sequence proposed by [23] too, and thus, a choice between the two methods will be mainly based on observed empirical results.

In summary, in order to solve a problem of the form (2) using the APG algorithm, one needs to identify or upper-bound $\|B\|_{\mathcal{L}(\mathcal{X},\mathcal{G})}$, find an expression for B^* and ∇g, identify β, and be able to obtain $\text{prox}_{\gamma f}(x)$ for any $x \in \mathcal{X}$.

B. Accelerated Proximal Gradient for Weighted Group-Sparse-Regularized Inverse Diffusion

In this section, we introduce the results that allow us to solve (1) using the APG algorithm. First, note that (1) is of the form (2) by identifying, with respect to the notation in Section I-A, the Hilbert spaces $\mathcal{G} = \mathcal{D}$, $\mathcal{X} = \mathcal{A}$, the operator $B = A : \mathcal{A} \rightarrow \mathcal{D}$, the lower semi-continuous non-smooth convex function $f : \mathcal{A} \rightarrow \bar{\mathbb{R}}$ such that

$$f(a) = \delta_{\mathcal{A}_+}(a) + \lambda \left\| \left\| \xi a_{\mathbf{r}} \right\|_{\mathrm{L}^2(\mathbb{R}_+)} \right\|_{\mathrm{L}^1(\mathbb{R}^2)}, \tag{4}$$

$\forall a \in \mathcal{A}$, and the lower semi-continuous, Frchet-differentiable convex function $g : \mathcal{D} \rightarrow \mathbb{R}$ such that

$$g(d) = \|d - d_{\text{obs}}\|^2_{\mathcal{D}}, \forall d \in \mathcal{D}. \tag{5}$$

Consequently, to derive the APG algorithm to solve (1), we use some of the results we obtained in Part I [19, Section III-B] on the diffusion operator A, namely, a bound on its norm and the expression for its adjoint. Furthermore, we need to find a $\beta > 0$ such that ∇g is β^{-1}-Lipschitz continuous, and provide a way to compute $\text{prox}_{\gamma f}(a)$ for any $\gamma > 0$ and $a \in \mathcal{A}$.

We start by characterizing the behavior of the smooth function g in (5). The result in Property 1 follows finite-dimensional intuition and specifies this behavior completely.

Property 1: (Fréchet derivative of the squared-norm loss). Consider the functional $g : \mathcal{D} \rightarrow \mathbb{R}$ in (5). Then, g has a Frchet derivative $\nabla g : \mathcal{D} \rightarrow \mathcal{D}^*$ such that $\nabla g(d)$ is represented in $\mathcal{D}$ by $2(d - d_{\text{obs}})$, $\forall d \in \mathcal{D}$. Additionally, $\forall d_1, d_2 \in \mathcal{D}$

$$\|\nabla g(d_1) - \nabla g(d_2)\|_{\mathcal{D}^*} = \|2d_1 - 2d_2\|_{\mathcal{D}} = 2\,\|d_1 - d_2\|_{\mathcal{D}}$$

and, thus, ∇g is 2-Lipschitz continuous, i.e., $\beta = 1/2$. Here, $\mathcal{D}^*$ is the dual space of $\mathcal{D}$, where $\nabla g(d)$ resides. See [37] for more on dual spaces and Frchet derivatives.

Now, we turn our attention towards the non-smooth function f in (4). Deriving a closed form expression for $\text{prox}_{\gamma f}(a)$ is the most complex result required to use the APG algorithm. This is mainly because the proximal operator does not generally decompose well for sums of functions, and no previous result indicates that $\text{prox}_{\gamma f}(a)$ can be computed in closed form. The appendix of this paper is dedicated to proving our contribution in Theorem 1, which provides an expression for $\text{prox}_{\gamma f}$ in the most generic setting possible. To simplify the exposition of this result, let $\aleph = \text{supp}\left(\xi\right)$ and $\aleph^{\mathsf{c}} = [0, \sigma_{\max}] \setminus \aleph$. These two sets distinguish values of σ at which the recovered PSDR a in (1) is influenced by the weighted group-sparsity regularization, i.e., $\sigma \in \aleph$, from values of σ at which it is not, i.e., $\sigma \in \aleph^{\mathsf{c}}$. Consider also, for any $a \in \mathcal{A}$, two functions $a_{\aleph}, a_{\aleph^{\mathsf{c}}} : \Omega \rightarrow \mathbb{R}_+$ such that

$$\text{supp}\left(a_{\aleph}\right) \subset \mathbb{R}^2 \times \aleph,\ \text{supp}\left(a_{\aleph^{\mathsf{c}}}\right) \subset \mathbb{R}^2 \times \aleph^{\mathsf{c}},$$

and $a = a_{\aleph} + a_{\aleph^{\mathsf{c}}}$, which provides a way for us to refer to these two distinct regions of the PSDR a.

Theorem 1 (Proximal operator of the non-negative weighted group-sparsity regularizer): Consider the functional $f : \mathcal{A} \rightarrow \bar{\mathbb{R}}$ in (4). For any $\gamma, \lambda > 0$, if $p = \text{prox}_{\gamma f}(a)$, then,

$$p_{\mathbf{r}} = [a_{\mathbf{r}}]_+ - \mathrm{P}_{\bar{\mathcal{B}}_{\xi}(\lambda\gamma)}\left[[a_{\aleph,\mathbf{r}}]_+\right].$$

Here, $\mathrm{P}_{\bar{\mathcal{B}}_{\xi}(\lambda\gamma)}$ is the projection onto $\bar{\mathcal{B}}_{\xi}\left(\lambda\gamma\right)$, the closed ellipsoid of ξ^{-1}-weighted norm under $\lambda\gamma$. This convex set and the projection onto it are further discussed in the appendix. Following the convention used in (1), for each $\mathbf{r} \in \mathbb{R}^2$, we have $p_{\mathbf{r}}, a_{\aleph,\mathbf{r}} : [0, \sigma_{\max}] \rightarrow \mathbb{R}_+$ such that $a_{\aleph,\mathbf{r}}(\sigma) = a_{\aleph}(\mathbf{r}, \sigma)$ and $p_{\mathbf{r}}(\sigma) = p(\mathbf{r}, \sigma)$ for any $\sigma \in [0, \sigma_{\max}]$.

The interpretation of this result is a direct parallel with the interpretation of the FISTA iterations in the known framework of ℓ^1-regularized inverse problems, i.e., an iterative shrinkage-thresholding effect. To see this, consider a specific $\mathbf{r} \in \mathbb{R}^2$ at which $w(\mathbf{r}) > 0$, and analyze the iteration of Step 4 in Algorithm 1. The iteration of the proximal operator in Theorem 1 will keep shrinking $[a_{\aleph,\mathbf{r}}]_+$ by subtracting its projection onto the ellipsoid $\bar{\mathcal{B}}_\xi(\lambda\gamma)$. If the gradient step inside does not raise this contribution again due to its importance for the minimization of g, at some point we will have $[a_{\aleph,\mathbf{r}}]_+ \in \bar{\mathcal{B}}_\xi(\lambda\gamma)$, which will result in a thresholding effect, because then applying the proximal operator in Theorem 1 again will yield $[a_{\aleph,\mathbf{r}}]_+ = 0$. In this context, we can read Theorem 1 as a statement that the non-negativity constraint and the weighted norm in (4) decouple, neither affecting the optimality of iterative shrinkage-thresholding for inducing sparsity.

Expressing the projection in Theorem 1 in closed form for a generic ξ and for each $\mathbf{r} \in \mathbb{R}^2$, however, is not trivial. In Property 2 in the appendix, we generalize a well-known finite dimensional result that states that this projection can not generally be fully determined in closed form and, thus, iterative procedures should be used for each $\mathbf{r} \in \mathbb{R}^2$. Although this establishes an interesting research direction to obtain algorithms that solve (1) in its more generic form, we opt here for limiting our choice of ξ. In particular, we select only its support $\aleph$ and we let $\xi = 1$ a.e. in $\aleph$. This simplifies the projection onto the closed ellipsoid $\bar{\mathcal{B}}_\xi(\lambda\gamma)$, which becomes the simple closed ball of norm smaller than $\lambda\gamma$ in $\mathrm{L}^2(\aleph)$ (see Property 3 in the appendix). For this particular case, Theorem 2 states a closed-form expression for $\operatorname{prox}_{\gamma f}(a), \forall a \in \mathcal{A}$, completing the list of required results to use the APG algorithm.

Theorem 2 (Proximal operator of the non-negative group-sparsity regularizer on $\aleph$): Consider the functional $f : \mathcal{A} \to \bar{\mathbb{R}}$ in (4). Let $\xi = 1$ a.e. in $\aleph$. Then, $\forall \gamma, \lambda > 0$, if $p = \operatorname{prox}_{\gamma f}(a)$,

$$p_{\mathbf{r}} = [a_{\aleph^c,\mathbf{r}}]_+ + [a_{\aleph,\mathbf{r}}]_+ \left(1 - \frac{\gamma\lambda}{\left\|[a_{\aleph,\mathbf{r}}]_+\right\|_{\mathrm{L}^2(\aleph)}}\right)_+ .$$

Here, $a_{\aleph,\mathbf{r}}$ and $p_{\mathbf{r}}$ are defined as in Theorem 1 and $a_{\aleph^c,\mathbf{r}}$ is defined mutatis mutandis.

This result, jointly with the bound on the diffusion operator's norm derived in Part I [19, Section III-B], is summarized in the proposed algorithm for inverse diffusion, i.e., Algorithm 2. This algorithm establishes a reference from which different discretization schemes can lead to different implementable algorithms for inverse diffusion and cell detection. A relevant observation here is that, precisely because $\operatorname{prox}_{\gamma f} = \operatorname{prox}_{\gamma\lambda f_1} \circ \mathrm{P}_{\mathcal{A}_+}$, where f_1 is the group-sparsity regularizer as in (1), the implementation of $\operatorname{prox}_{\gamma f}$ is decomposed in the non-negative projection in Step 4 and the subsequent group-sparsity shrinkage-thresholding in Step 5.

Algorithm 2: Accelerated Proximal Gradient to find a_{opt} that solves (1) with function-value convergence rate $\mathcal{O}\left(i^{-2}\right)$ when $\xi = 1$ a.e. in $\aleph$. Here, $\eta = \sigma_{\max}^{-1} \|w\|_{\mathrm{L}^\infty(\mathbb{R}^2)}^{-2}$ is used for clarity of exposition.

```
Require: An initial $a^{(0)} \in \mathcal{A}$, an image observation
         $d_{\mathrm{obs}} \in \mathcal{D}$
Ensure: A solution $a_{\mathrm{opt}} \in \mathcal{A}$ that solves (1)
1: $b^{(0)} \leftarrow a^{(0)}, i \leftarrow 0$
2: repeat
3:     $i \leftarrow i + 1, \alpha \leftarrow \frac{t(i-1)-1}{t(i)}$
4:     $a^{(i)} \leftarrow \left[b^{(i-1)} - \eta A^* \left(A b^{(i-1)} - d_{\mathrm{obs}}\right)\right]_+$
5:     $a_{\aleph}^{(i)} \leftarrow a_{\aleph}^{(i)} \left(1 - \frac{\eta}{2}\lambda \left\|a_{\mathbf{r},\aleph}^{(i)}\right\|_{\mathrm{L}^2(\aleph)}^{-1}\right)_+$
6:     $b^{(i)} \leftarrow a^{(i)} + \alpha\left(a^{(i)} - a^{(i-1)}\right)$
7: until convergence
8: $a_{\mathrm{opt}} \leftarrow a^{(i)}$
```

C. Discretization of the APG for Inverse Diffusion

In Part I [19, Section IV] we presented a discretization scheme that establishes approximation rules for any element in $\mathcal{A}$ by an element of $\mathbb{T}(M, N, K)$, and for any element in $\mathcal{D}$ by an element of $\mathbb{T}(M, N)$. Here, M and N are the number of pixels in each dimension, and K is the number of discretization points for the σ-dimension. This discretization scheme also enables us to obtain discrete versions of the diffusion operator A and its adjoint A^*, and yields Algorithm 3 as a practical implementation of Algorithm 2. In Algorithm 3, $\tilde{d}_{\mathrm{obs}}$, $\tilde{w}$, $\tilde{\mu}$ and $\tilde{a}$ are discretizations of d_{obs}, w, μ and a, $\tilde{\aleph}$ is the set of indexes k that represent portions of the σ-dimension that lie inside $\aleph$, and $\tilde{g}_k$ are the doubly spatially integrated Gaussian kernels, as specified in Part I [19, Section IV]. In Algorithm 3, Steps 1, 3, and 12 take care of the Nesterov acceleration of the proximal gradient algorithm by using the momentum in its convergence path, Step 4 computes the diffusion operator and evaluates the prediction error, Step 6 computes the adjoint operator, completes the gradient step, and enforces the positivity constraint, and Steps 8 and 10 implement the group-sparsity shrinkage-thresholding.

Many of the choices implicit in the discretization scheme presented in Part I [19, Section IV] were derived from an intuitive goal, i.e., that the properties present in the function spaces are preserved after discretization. In this manner, the discretized adjoint is the adjoint of the discretized operator, and the proximal operator is preserved, because the discretized and continuous norms are equivalent in an inner-approximation sense. As a result, Algorithm 3 is an APG algorithm too, and it can be proven to solve the discretized equivalent to (1), i.e., (6), for $\tilde{a} \in \mathbb{T}_+(M, N, K)$ (see Part I [19, Equation (24)]).

$$\min_{\tilde{a}} \left\{ \left\| \tilde{d}_{\mathrm{obs}} - \sum_{k=1}^{K} \tilde{g}_k \circledast \tilde{a}_k \right\|_{\tilde{w}}^2 + \lambda \sum_{m,n} \sqrt{\sum_{k \in \tilde{\aleph}} \tilde{a}_{m,n,k}^2} \right\} \quad (6)$$

III. Numerical Results

In this section, we provide empirical validation of the optimization framework we presented in Part I [19, Section III], i.e., (1), and of the theoretical results in Section II. We do this through the evaluation of an efficient approximated implementation of Algorithm 3 we present in Section III-A. In particular, in Section III-B we specify how we use the observation model we presented in Part I [19, Section II-B] to generate realistic synthetic data, in which the location and total secretion of each

Algorithm 3: Algorithm to find a discrete approximation $\tilde{a}_{\mathrm{opt}} \in \mathbb{T}\left(M, N, K\right)$ to the solution of (1), i.e., the solution to (6). Here, η is as in Algorithm 2, $\circledast$ refers to discrete zero-padded same-size convolution, and all matrix powers and products are element-wise.

```
Require: An initial $\tilde{a}^{(0)} \in \mathbb{T}\left(M, N, K\right)$, a discrete image
         observation $\tilde{d}_{\mathrm{obs}} \in \mathbb{T}\left(M, N\right)$
Ensure:  A discrete approximation $\tilde{a}_{\mathrm{opt}} \in \mathbb{T}\left(M, N, K\right)$ to
         the solution of (1), i.e., the solution to (6)
 1: $\tilde{b}^{(0)} \leftarrow \tilde{a}^{(0)}$, $i \leftarrow 0$
 2: repeat
 3:     $i \leftarrow i+1$, $\alpha \leftarrow \frac{t(i-1)-1}{t(i)}$
 4:     $\tilde{d}^{(i)} \leftarrow \sum_{k=1}^{K} \tilde{g}_k \circledast \tilde{b}_k^{(i-1)} - \tilde{d}_{\mathrm{obs}}$
 5:     for $k = 1$ to $K$ do
 6:     $\tilde{a}_k^{(i)} \leftarrow \left[\tilde{b}_k^{(i-1)} - \eta\tilde{\mu} \odot \left(\tilde{g}_k \circledast \left[\tilde{w}^2 \odot \tilde{d}^{(i)}\right]\right)\right]_+$
 7:     end for
 8:     $\tilde{p} \leftarrow \left(1 - \frac{\eta}{2}\lambda \left[\sqrt{\sum_{k \in \tilde{\aleph}} \left(\tilde{a}_k^{(i)}\right)^2}\right]^{-1}\right)_+$
 9:     for $k \in \tilde{\aleph}$ do
10:         $\tilde{a}_k^{(i)} \leftarrow \tilde{p} \odot \tilde{a}_k^{(i)}$
11:     end for
12:     $\tilde{b}^{(i)} \leftarrow \tilde{a}^{(i)} + \alpha\left(\tilde{a}^{(i)} - \tilde{a}^{(i-1)}\right)$
13: until convergence
14: $\tilde{a}_{\mathrm{opt}} \leftarrow \tilde{a}^{(i)}$
```

TABLE I

PARAMETERS USED FOR DATA-GENERATION IN OUR SIMULATIONS. NOTE THAT THE SELECTION OF THESE PARAMETERS HAS BEEN DONE WITHIN REALISTIC RANGES (SEE, AMONG OTHERS, [39] FOR ELISPOT EXAMPLES). NONETHELESS, κ_{a} AND σ_{b} HAVE BEEN SPECIFICALLY ADJUSTED TO MATCH THE ASPECT OF THE REAL FLUOROSPOT DATA THAT WAS AVAILABLE. FINALLY, NOTE THAT WITH THESE PHYSICAL PARAMETERS, $\tilde{\sigma}_{\mathrm{max}} = \sqrt{2DT}/\Delta_{\mathrm{pix}} \approx 64.5$

κ_{a} $\left[\mathrm{ms}^{-1}\right]$	κ_{d} $\left[\mathrm{s}^{-1}\right]$	D $\left[\mathrm{m}^2\mathrm{s}^{-1}\right]$	T [h]
10^{-7}	10^{-4}	$3 \cdot 10^{-12}$	8

Δ_{pix} [μm]	M	N	K_{g}	J	N_t	$\tilde{\sigma}_{\mathrm{b}}$ [pix]
6.45	512	512	30	10	10^3	2.28

of the active cells is known. On that data, we evaluate our approach in two different ways. First, in Section III-C1, we provide detection performance metrics after simple post-processing, and compare that to the detection performance of a human expert on similarly generated data. This, jointly with our results on real data in Part I [19, Section V-A], validates our proposal for use in practical scenarios. Second, in Section III-C2, we evaluate the output $\tilde{a}_{\mathrm{opt}}$ of Algorithm 3 by interpreting its accumulated sum over k as a 2D discrete particle distribution, and comparing it to the one given by the true simulated PSDR $\tilde{a}$ using optimal-transport theory.

A. Implementation, Kernel Approximations

The main driving factor of the computational cost of Algorithm 3 is the $2K$ convolutions with 2D kernels $\tilde{g}_k$ at each iteration in Steps 4 and 6. Although efficiently parallelizable in GPUs, 2D convolution is still an expensive operation. Recall from Part I [19, Section IV] that the discretized filters $\tilde{g}_k$ are given by

$$\tilde{g}_k[(m,n)] = \frac{1}{\sqrt{\Delta_k}} \int_{\tilde{\sigma}_{k-1}}^{\tilde{\sigma}_k} \omega_{\tilde{\sigma}}(m)\omega_{\tilde{\sigma}}(n)\mathrm{d}\tilde{\sigma}\,, \tag{7}$$

for some $\omega_{\tilde{\sigma}} : \mathbb{Z} \to \mathbb{R}_+$, where $\Delta_k = \tilde{\sigma}_k - \tilde{\sigma}_{k-1}$ is the width of the $\tilde{\sigma}$-dimension bin represented by k. This expression suggests that $\tilde{g}_k$ is close to being separable, at least for small values of Δ_k. In the particular choice of parameters for our analysis on synthetic data, detailed in Section III-B, Table III, the smallest value of $\lambda_1 / \sum \lambda_l$, where λ_l are the decreasingly sorted singular values of a given kernel $\tilde{g}_k$, was 97.72 %, while the smallest value of $(\lambda_1 + \lambda_2 + \lambda_3)/\sum \lambda_l$ was 99.99 %. In this context, we propose to approximate the 2D kernels $\tilde{g}_k$ by separable 2D kernels, i.e., rank-one kernels. Thus, we will approximate each convolution with a 2D kernel by 2 successive convolutions with 1D kernels, substantially reducing the computational effort. Note, however, that regardless the approximation, the $2K$ convolutions per iteration will still remain the bottleneck of Algorithm 3, and thus, further efforts on the reduction of the computational burden should involve efficient techniques to implement these convolutions.

In Section III-C, we report the results of approximating $\tilde{g}_k$ as g_k^{br1}, the best rank-one approximation in terms of the Frobenius norm. g_k^{br1} can be obtained numerically by using singular value decomposition on the original kernel $\tilde{g}_k$. In the supplementary material to this paper, we discuss two simpler rank-one approximations and report their performance, which was significantly worse than that of g_k^{br1} in almost every scenario. In order to quantify the loss in performance due to the rank-one approximation, we will also include in Section III-C the results using g_k^{br3}, the best rank-three approximation in terms of the Frobenius norm, which approximates every convolution with a 2D kernel by combining 6 convolutions with 1D kernels.

B. Data Simulation

We simulated image data from a physical system that follows the reaction-diffusion-adsorption-desorption process we presented in Part I [19, Section II-A] with the parameters specified in Table I. Here, we have that 1) κ_{a}, κ_{d}, D and T are physical parameters characterizing the biochemical assays, 2) M, N and Δ_{pix} determine the spatial discretization of a supposed camera, as detailed in Part I [19, Section IV], 3) N_t determines the number of discretization points in time used to generate the SDR $s(\mathbf{r}, t)$, 4) K_{g} determines the number of uniform discretization intervals of the PSDR $a(\mathbf{r}, \sigma)$ in the σ-dimension during data generation, 5) J determines the number of terms to which we truncate the infinite sum that expresses $\varphi(\tau, t)$, the function that relates the SDR s to the PSDR a, as we exposed in Part I [19, Section II-C, Lemma 2], and 6) $\tilde{\sigma}_{\mathrm{b}}$ determines the standard deviation of the discretized Gaussian kernel used to simulate an imperfect optical system, as presented in Part I [19, Section II-D].

TABLE II
CHARACTERISTIC PARAMETERS OF THE 12 DIFFERENT SCENARIOS CONSIDERED IN THE SIMULATIONS, FORMED BY FOUR NOISE LEVELS (NL) AND THREE CELL DENSITIES

	Few	Average	Many
N_c	250	750	1250

	NL 1	NL 2	NL 3	NL 4
b	10	8	6	4

TABLE III
PARAMETERS USED FOR ALGORITHM 3 IN ALL OF THE SIMULATIONS PRESENTED IN THIS PAPER. NOTE HERE THAT THE CHOICE OF THE GRID IN THE $\tilde{\sigma}$-DIMENSION IS COHERENT WITH THE OBSERVATION MODEL UNDER AN IMPERFECT OPTICAL SYSTEM DERIVED IN [19, SECTION II-D], i.e., WITH RESPECT TO TABLE I, $\tilde{\sigma}_0 \approx \tilde{\sigma}_\mathrm{b}$ AND $\tilde{\sigma}_K \approx \tilde{\sigma}_{\max} + \tilde{\sigma}_\mathrm{b}$

K	$\tilde{\sigma}_0, \tilde{\sigma}_1, \ldots, \tilde{\sigma}_8$	I	$\tilde{\aleph}$
8	$2.3, 5, 9, 13, 23, 33, 43, 53, 67$	10^4	$\{1, 2, \ldots, 8\}$

For each considered active cell, say, in a location (m, n), we generated a random discrete SDR $\tilde{s}_{m,n}$ in the form of a square pulse in time. In particular, we drew uniform activation (particle generation initiation) and deactivation (particle generation finalization) times in the interval $(1, 6)$h, and we chose the amplitude of the square pulse by uniformly drawing the total amount of generated particles between a certain maximum and its half. This was done for 50 different sets of uniformly-drawn, pixel-centered, active-cell locations for each considered number N_c of active cells in an image.

We then used our contribution in Part I [19, Theorem 2] to obtain the PSDR $\tilde{a} \in \mathbb{T}\left(M, N, K_\mathrm{g}\right)$ from the resulting SDR $\tilde{s} \in \mathbb{T}\left(M, N, N_t\right)$. Details on the exact procedure to do so can be found in the supplementary material to this paper. Then, we computed the ideal discretized measurement by applying the discretized diffusion operator $\tilde{A}$ to it. Note here that in synthesis, the kernels $\tilde{g}_k$ were not approximated. We then simulated the effect of an imperfect optical system by convolution with a discretized, i.e. spatially integrated, Gaussian kernel with a standard deviation of $\tilde{\sigma}_\mathrm{b}$, and rescaled the image to keep the intensity in the range $[0, 1]$. We then incorporated additive white Gaussian noise of the variance that corresponded to that of the statistical model for quantization in the range $[0, 1]$ with a number of bits b, i.e. $2^{-2b}/12$. Finally, we clipped the resulting image to the range $[0, 1]$ and re-scaled it to the range $[0, 255]$. It is worthwhile to mention here that extensive analysis carried out on real data has suggested that the Gaussian assumption is sensible. Moreover, no magnification is usually employed in the image capture for the described biochemical assays. This implies high photon counts, which theoretically supports the Gaussian assumption over the Poisson assumption, more common in low-photon-count applications such as microscopy.

Throughout this section, we will present results obtained by this data-generation procedure in twelve different scenarios, in which three different cell densities (few, average, many) and four different noise levels (NL) are considered. For details on their characterization in terms of N_c and b, see Table II.

C. *Performance Evaluation and Numerical Results*

The empirical evaluation of Algorithm 3 can be addressed in terms of diverse metrics. On one hand, one could focus on metrics characteristic of the optimization framework itself, i.e., the prediction's square error, the group-sparsity level in the solution, or the value of the cost function from (1) and the rate at which it decreases. Fig. 1 exemplifies the statistics of these quantities during convergence. These metrics, however, have already been studied theoretically and do not hold operational meaning in terms of performance on the task at hand, i.e., SL on data from reaction-diffusion-adsorption-desorption systems. On the other hand, detection metrics such as precision and recall, or their compromise, the F1-score, directly characterize SL performance, and are therefore naturally operational. Therefore, when presenting results to validate the operational value of our algoriths, we will use the F1-score after $I = 10^4$ iterations, relying on convergence. For example, in Fig. 3 we compare our algorithm's F1-Score to that of an expert human labeler on synthetic data for some specific experimental conditions. Nonetheless, pure detection metrics like the F1-score can not be obtained simply from the value $\tilde{a}_\mathrm{opt}$ our algorithm provides, and some post-processing is necessary. Therefore, any attempt at evaluating our approach in this manner will be influenced by the specificities of the chosen post-processing. In this context, optimal transport theory and the earth mover's distance (EMD) [40] offer an interesting alternative. In particular, the EMD is an interpretable objective metric between any two discrete distributions of the same total weight. In other words, it not only evaluates the location at which each spatial peak in the recovered $\tilde{a}_\mathrm{opt}$ is, but also their relative contribution to the total amount of particles. In conclusion, then, when evaluating our results in terms of the accuracy of the information they provide about the spatial distribution of particle generation, we will use the EMD as our preferred metric.

1) Operational Evaluation and Detection Results: Consider an SL detector that, given an observation $\tilde{d}_\mathrm{obs}$, provides a list of positions $\{\mathbf{r}_l\}_{l=1}^{L} \subset \mathbb{R}^2$ and a co-indexed list of positive numbers (pseudo-likelihoods) $\{p_l\}_{l=1}^{L} \subset \mathbb{R}_+$. Then, for a given tolerance $\varrho > 0$, we will evaluate each position $\mathbf{r}_l$ in decreasing order of pseudo-likelihood p_l, and consider it a correct detection if a previously unmatched true cell location $\mathbf{r}_\mathrm{c}$ can be found inside the ball with diameter ϱ centered at $\mathbf{r}_l$. If that is the case, the closest such true cell location will not be paired with any further $\mathbf{r}_l$s. Then, if we refer to TP, FP and FN as the number of correct detections, incorrect detections, and cells that were not detected, respectively, the precision pre, recall rec and F1-score F1 are defined as

$$\mathrm{pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \ \mathrm{rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \ \text{and } \mathrm{F1} = \frac{2\,\mathrm{pre} \cdot \mathrm{rec}}{\mathrm{pre} + \mathrm{rec}}.$$

Note, then, that the F1-score is a number in the range $[0, 1]$ that establishes a compromise between the probability of a detection being correct (precision) and the probability of a true cell being found (recall). Throughout the rest of the paper, we will use $\varrho = 3$ pix as our tolerance for the localization of active cells. Note here that, as mentioned in [19, Section II-A], the cells under

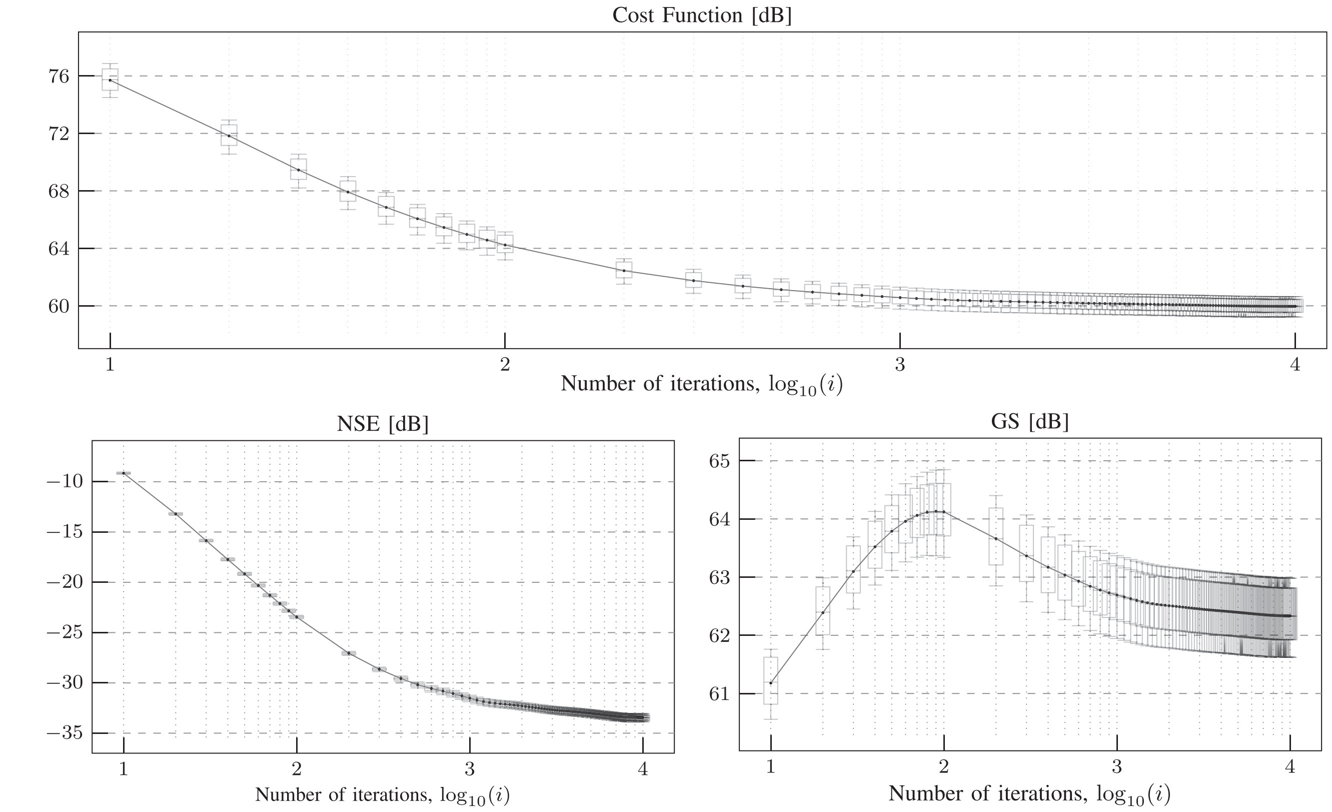


Fig. 1. Statistics of the optimization metrics' convergence with the number of iterations i. Showing the normalized prediction's square error $\text{NSE} = \|Aa - d_{\text{obs}}\|_{\mathcal{D}}^2 / \|d_{\text{obs}}\|_{\mathcal{D}}^2$, the value of the group sparsity regularizer (GS), and the value of the cost function in (1). Comprising results from 50 images with $N_{\text{c}} = 750$ cells and noise level 3 (see Table II) when analyzed with Algorithm 3 with the parameters in Table III and $\lambda = 0.5$. For each quantity, a dot and the line illustrate mean behavior, whiskers indicate the evolution of the 10th and 90th percentiles, and the box indicates the evolution of the 25th, 50th and 75th percentiles.

consideration are tens of μms in diameter, and so a tolerance of $\varrho\Delta_{\text{pix}} = 19.5\ \mu$m should be considered extremely accurate.

Obtaining a set of detections $\{(\mathbf{r}_l, p_l)\}_{l=1}^{L}$ from the output of Algorithm 3 can be done in multiple ways. In an ideal case, i.e., with the perfect reconstruction of $\tilde{a}$, we would simply use $\{\tilde{\mathbf{r}}_l\}_{l=1}^{L} = \bigcup_{k=1}^{K} \text{supp}\,(\tilde{a}_k)$, where the support of a matrix is the set of indexes $\tilde{\mathbf{r}} \in \mathbb{Z}^2$ where its elements are not zero. In this case, the value of p_l would not have any impact, and the obtained F1-score would be 1. In real cases, in which an imperfect reconstruction $\tilde{a}_{\text{opt}}$ includes approximation and numerical errors, we propose to first compute a pseudo-likelihood for each pixel, corresponding to the contribution of each pixel to the overall group-sparsity regularizer, i.e. the matrix $\tilde{p} = \left(\sum_{k\in\tilde{\aleph}} \tilde{a}_k^2\right)^{1/2}$. We then propose to build a list of candidate detections $\{\tilde{\mathbf{r}}_q\}_{q=1}^{Q}$ formed by the local maxima (with 8-connectivity) in $\tilde{p}$, and discard those with pseudo-likelihood $p_q = \tilde{p}_{\tilde{\mathbf{r}}_q}$ under a certain threshold, i.e., for some $\delta > 0$, $\{(\tilde{\mathbf{r}}_l, p_l)\}_{l=1}^{L} = \{(\tilde{\mathbf{r}}_q, p_q) : p_q > \delta\}$. In practice, we pick the δ that yields the best F1-score given the known true data. This mimics real application, in which experts select the threshold that best fits their criterion by visual inspection of the results overlaid on the image data. This same evaluation by connected maxima detection and optimal thresholding can be applied to other methods in which the pseudo-likelihood image $\tilde{p}$ is generated differently. Finally, note that although different heuristics could generate a better set of detections and pseudo-likelihoods $\{\mathbf{r}_l, p_l\}_{l=1}^{L}$, our focus here is in showing that the PSDR $\tilde{a}$ recovered from Algorithm 3 provides the means for robust and reliable SL.

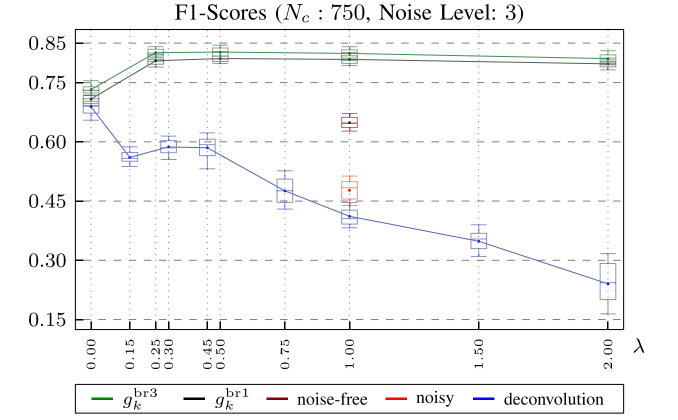


Fig. 2. Statistics of the obtained F1-scores for different methods to obtain $\tilde{p}$. Dependence on the regularization parameter λ. Those methods that do not use a regularization parameter appear centered in the figure. The statistics are reported in the whiskers-box plot as in Fig. 1.

In Figs. 2, 3, and 5, we report statistics on the results of using $\tilde{p}$ computed as above when Algorithm 3 is used with the parameters in Table III and the sequence $t : \mathbb{N} \to \mathbb{R}_+$ suggested in [23]

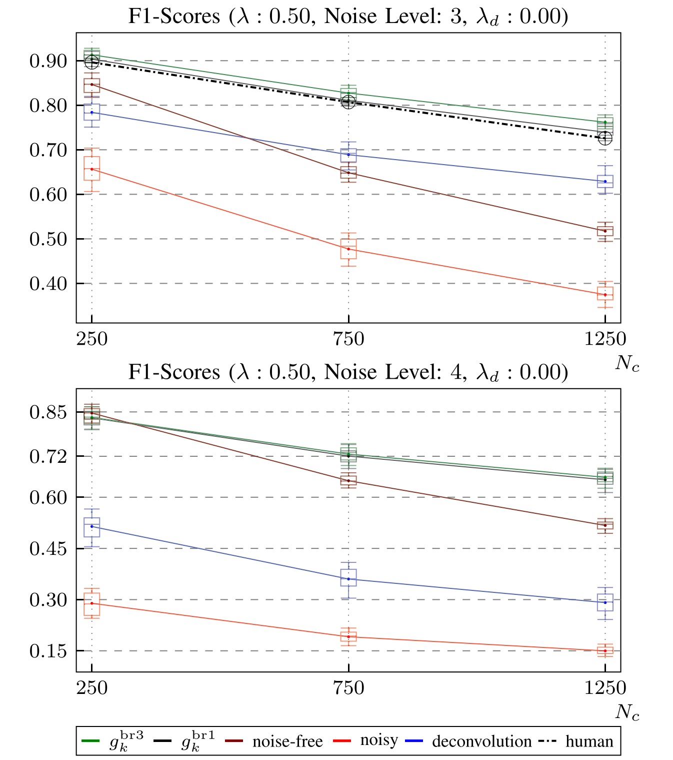


Fig. 3. Statistics of the obtained F1-scores for different methods. Dependence on the total number of cells in the simulated image, at the two highest noise levels. For noise level 3, performance obtained by an expert human labeler on one image of each density. Regularization parameters $\lambda = 0.5$ and $\lambda_{\mathrm{d}} = 0$ chosen for their respective methods due to the results in Fig. 2. The statistics are reported in the whiskers-box plot as in Fig. 1.

(see Section II-A for details). Note here that $K = 8$ implies that the discretization in the $\tilde{\sigma}$-dimension used in the analysis is much coarser than that used in the generation of the data, i.e., $K_{\mathrm{g}} = 30$. Note also that due to the large amount of decisions involved in choosing the discretization of the $\tilde{\sigma}$-dimension, this was done manually by trial-and-error and always maintaining SL performance in mind. In this sense, the lowest $\tilde{\sigma}$ s were discretized with more detail, since they allow for a more accurate localization of the active cells' position. Finally, note that $\tilde{w}(\tilde{\mathbf{r}}) = 1$ and $\tilde{\mu}(\tilde{\mathbf{r}}) = 1$ were used in the context of the simulated data.

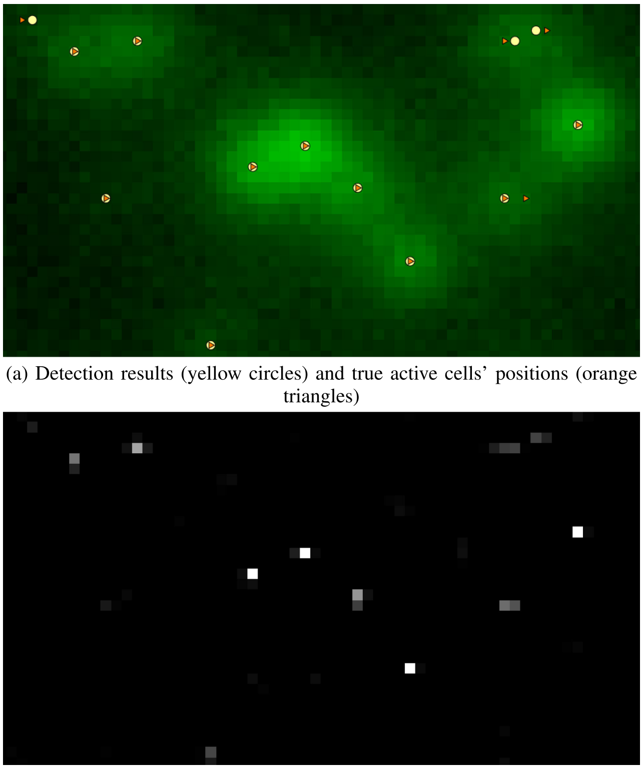
To provide a fair evaluation, we compare the obtained results with different proposals for $\tilde{p}$. As a baseline for comparison, we obtain the results of using a noise-free version of the observed image $\tilde{d}_{\mathrm{obs}}$ as $\tilde{p}$. Because under the observation model $Aa = d_{\mathrm{obs}}$, isolated active cells lead to monomode profiles in d_{obs} around the true cell location, this will provide a reference on how detection is affected by interactions between different active cells. At the same time, this will also provide an upper bound on the performance of any denoising-centered approach. Similarly, we will also obtain the results of using $\tilde{p} = \tilde{d}_{\mathrm{obs}}$ directly, which will provide a reference on how detection is affected by additive noise. Finally, we also provide the results of obtaining $\tilde{p}$ from a sparsity-based deconvolution scheme on $\tilde{d}_{\mathrm{obs}}$ that aims to invert the blur introduced by the optics. This latter approach is implemented by using $I = 10^4$ iterations of Algorithm 3 with $K = 1$ and g_1 the same kernel used to simulate the optical imperfections, and with the step-length η optimized to obtain the empirically best results, in terms of both performance and robustness, i.e., $\eta \approx 0.44$. For the analysis of Figs. 2, 3, and 5, we will consider that the difference between two quantities is statistically significant if the 10th empirical percentile of one of the two quantities is above the 90th empirical percentile of the other.

Both the sparsity-based deconvolution scheme and our own approach rely on an hyper-parameter λ that needs to be selected. As is common in sparsity-based optimization frameworks, this choice is made here experimentally. Fig. 2 shows the statistics of the F1-score for the considered methods as a function of λ for $N_{\mathrm{c}} = 750$ and the third noise level considered in Table II. Methods that do not depend on λ are additionally reported for comparison.

Fig. 2 suggests that the choice of regularizer in the optimization framework (1) proposed in Part I [19, Section III] is beneficial for SL. Indeed, regardless of the approximation used, any of the tested values for the regularization parameter λ yield significantly better F1-scores than $\lambda = 0$. Furthermore, the results in Fig. 2 indicate that (1) is robust to the choice of regularization parameter λ, showing practically no change in performance across a whole order of magnitude, i.e., from $\lambda = 0.15$ to $\lambda = 2$. In contrast, the results for sparsity-based deconvolution indicate that ℓ^1-regularization is not appropriate in this setting, and that, if used, the choice of regularization parameter will be critical to the obtained performance. Additionally, Fig. 2 also validates our rank-one approximation strategy, as using g_k^{br3} yields only non-significant improvements on the performance obtained by using g_k^{br1} while triplicating the computational cost.

These conclusions, i.e., 1) that the regularizer chosen in (1) is adequate for inverse diffusion for SL, 2) that the proposed optimization framework is robust to the choice of regularization parameter, and 3) that the differences in performance when using a rank-one and a rank-three approximation of the kernels are non-significant, are preserved throughout the remaining eleven scenarios characterized by combinations of the parameters in Table II. Replicates of Fig. 2 for all possible combinations are reported in the supplementary material. Finally, using Fig. 2, we decide that for the remainder of our analysis we will use $\lambda = 0.5$ for our approach and $\lambda_{\mathrm{d}} = 0$ for deconvolution.

Of the two factors under consideration that affect SL performance, interference between several sources seems to be the hardest to address. Indeed, in Fig. 3 we see that all considered methods, including an expert human labeler, decay steeply in detection performance when dealing with higher densities of active cells. This is to be expected, because the closer any two active cells are, the more indistinguishable they will be on the spots that result in the observed image. In particular, we observe that the performance of the expert human labeler decays with N_{c} at a similar or at a steeper rate as that obtained by our methodology. This seems to indicate that there is a common limiting factor to these performances, to which our methodology is at least as robust as a domain expert. Further, Fig. 3 indicates that, for the tested cell densities, the human labeler consistently performs within the 10th and the 90th percentile of the results obtained by our approximated implementation of Algorithm 3, exhibiting no significant differences. Nonetheless, the gap between the performance obtained by using g_k^{br3} in



(a) Detection results (yellow circles) and true active cells' positions (orange triangles)

(b) $\tilde{p} = \sqrt{\sum_{k \in \tilde{\mathbb{N}}} \tilde{a}_k^2}$ obtained from Alg. 3 with the parameters in Tab. 7, $\lambda = 0.5$, and using g_k^{br1}, at increased luminosity.

Fig. 4. Example of SL performance on a section of a simulated image with $N_{\mathrm{c}} = 1250$ and noise level 4 (see Table II).

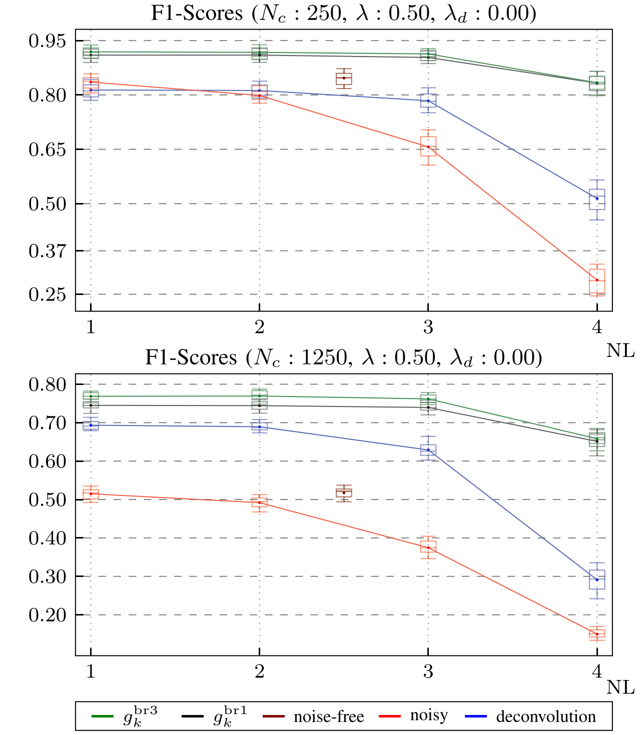


Fig. 5. Statistics of the obtained F1-scores for different methods. Dependence on the noise level, at the lowest and highest cell densities. The method that does not depend on noise appears centered in the figure. Regularization parameters' choice and reporting of statistics consistent with Fig. 2.

Algorithm 3 and that obtained by the expert human labeler does clearly increase with N_{c}, which suggests that substantial differences could have been observed at higher cell densities, i.e., for $N_{\mathrm{c}} > 1250$. Note here that due to the considerable amount of time required to manually label each image, only three synthetic images were manually labeled by the human expert, one of each cell density and at noise level 3. Further, note that the expert human labeler was unaware of the total expected number of cells in each image.

Fig. 3 also shows that Algorithm 3 provides significantly better SL performance than even picking local maxima in a noise-free version of the image. Indeed, only when our approach is exposed to a noise level 4 and we consider the lowest cell density does a noise-free image yield similar performances. This suggests that our approach is capable, through a noisy observation of $\tilde{d}_{\mathrm{obs}}$, of breaking apart clusters that would not exhibit local maxima at the active cells' locations even in a noise-free observation.

An example of this capacity of breaking clusters that do not exhibit maxima at the sought locations is illustrated in Fig. 4, where a section of a simulated image in the worst considered scenario (noise level 4, $N_{\mathrm{c}} = 1250$) is shown, together with its true active cell locations and the obtained detections.

In conclusion then, although Algorithm 3 is still incapable of telling apart cells that are arbitrarily close, it is well equipped to accurately detect active cells from spots generated by their combined secretion. In fact, Fig. 3 suggests that better approximations of the kernels g_k yield increased robustness in this sense, vouching for the proposed optimization framework [19, Section III], i.e., (1). In Fig. 4, note that most of the correctly detected cells are detected in the exact same pixel they were located, and all others are at a distance of one single pixel. This accuracy of the obtained locations has been observed consistently throughout our experimentation.

Finally, Fig. 3 also reveals a great advantage of Algorithm 3, i.e., robustness to additive noise. Indeed, Fig. 5 confirms that Algorithm 3, and our optimization framework exhibit an unparalleled robustness to additive noise, regardless the considered cell density.

2) Distributional Evaluation and Results: Consider now $\tilde{p} = \sum_{k=1}^{K_{\mathrm{g}}} \sqrt{\sigma_{\max}/K_{\mathrm{g}}}\tilde{a}_k$, the true spatial distribution of released particles within the discretization scheme of [19, Section IV], approximating the term $\int_0^{\sigma_{\max}} a d\sigma$. The ultimate objective of any source localization and characterization technique is to recover $\tilde{p}$ perfectly. Indeed, if one obtains $\tilde{p}$, one knows exactly how many particles were released from each location, and thus, the exact location of each source and their relative importance. For the evaluation of source localization and characterization algorithms, then, it is natural to consider whether an interpretable metric between $\tilde{p}$ and a recovered or estimated spatial density $\hat{p}$ is available.

The EMD [40] plays this role when two discrete distributions have the same total weight, i.e. if $\sum_{m,n=1}^{M,N} \tilde{p}_{m,n} = \sum_{m,n=1}^{M,N} \hat{p}_{m,n}$. In the following, we will consider this condition to be verified, and practically normalize both $\tilde{p}$ and $\hat{p}$ to

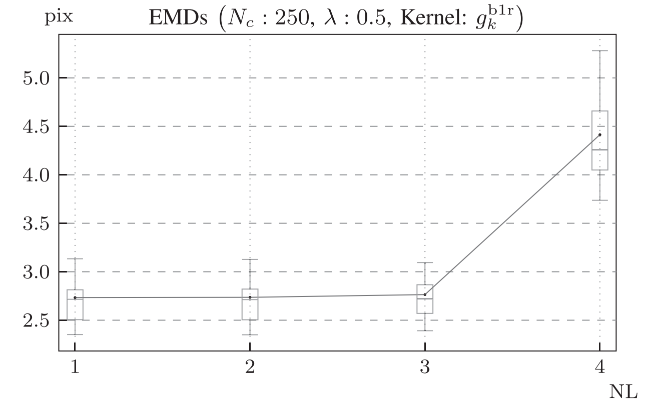


Fig. 6. Statistics of the EMD between $\tilde{p}$ and $\hat{p}$ resulting from Algorithm 3 using the best rank-one approximation to the discretized kernels and the parameters in Table III. Estimated from 50 images with $N_{\mathrm{c}} = 250$ cells. Dependence on noise level.

the same overall weight, i.e., the true number of cells N_{c}. The EMD can be interpreted as the minimal average displacement required to transform the estimated distribution $\hat{p}$ into the real distribution $\tilde{p}$. To formalize how the EMD is computed, consider $\mathcal{I} = \operatorname{supp}(\hat{p}) \subset \mathbb{Z}^2$ and $\mathcal{J} = \operatorname{supp}(\tilde{p}) \subset \mathbb{Z}^2$, and consider the Euclidean distance (in pix) between any two locations $\mathbf{i} \in \mathcal{I}$ and $\mathbf{j} \in \mathcal{J}$, i.e., $\|\mathbf{i} - \mathbf{j}\|_2$. Then, the following linear program

$$\begin{aligned} \min_{\tilde{f} \in \mathbb{T}(|\mathcal{I}|,|\mathcal{J}|)} \quad & \sum_{\mathbf{i}\in\mathcal{I}} \sum_{\mathbf{j}\in\mathcal{J}} \tilde{f}_{\mathbf{i},\mathbf{j}} \, \|\mathbf{i} - \mathbf{j}\|_2 \\ \text{subject to} \quad & \tilde{f}_{\mathbf{i},\mathbf{j}} \geq 0, \forall (\mathbf{i},\mathbf{j}) \in \mathcal{I} \times \mathcal{J}, \sum_{\mathbf{j}\in\mathcal{J}} \tilde{f}_{\mathbf{i},\mathbf{j}} \leq \hat{p}_{\mathbf{i}}, \forall \mathbf{i} \in \mathcal{I} \\ & \sum_{\mathbf{i}\in\mathcal{I}} \tilde{f}_{\mathbf{i},\mathbf{j}} \leq \tilde{p}_{\mathbf{j}}, \forall \mathbf{j} \in \mathcal{J}, \text{ and } \sum_{\mathbf{i}\in\mathcal{I}} \sum_{\mathbf{j}\in\mathcal{J}} \tilde{f}_{\mathbf{i},\mathbf{j}} = N_{\mathrm{c}}, \end{aligned}$$

is known as Monge-Kantorovich transportation problem. In our context, it determines the density of particles $f_{\mathbf{i},\mathbf{j}}$ that has to be moved from each location $\mathbf{i}$ to each location $\mathbf{j}$ so that $\hat{p}$ becomes $\tilde{p}$ with the minimal amount of overall work, understood as the product between the densities of particles and the distances they have to be moved. As a result, one can derive the average distance the density of particles has been transported (in pix), i.e., the EMD, as

$$\mathrm{EMD} = \sum_{\mathbf{i}\in\mathcal{I}} \sum_{\mathbf{j}\in\mathcal{J}} \frac{\tilde{f}_{\mathbf{i},\mathbf{j}}}{N_{\mathrm{c}}} \, \|\mathbf{i} - \mathbf{j}\|_2 \ ,$$

for the $\tilde{f} \in \mathbb{T}(|\mathcal{I}|, |\mathcal{J}|)$ that solved the linear program above.

In our setting, we used $\hat{p} = \sum_{k=1}^{K} \sqrt{\Delta_k}\, \tilde{a}_{\mathrm{opt},k}$, where $\tilde{a}_{\mathrm{opt}}$ is the PSDR recovered from Algorithm 3 with the same configuration as in the previous section and $\lambda = 0.5$. We solved the transportation problem above using CVX [41] with the MOSEK [42] solver and reported the statistics of the obtained EMD s for 50 images with $N_{\mathrm{c}} = 250$ for each of the four different noise levels of Table II in Fig. 6. There, we can observe that for the three first noise levels, the EMD consistently stays below 3 pix. This result is remarkable, because it does not only include the displacement of the highest peaks of particle secretion density, but also any errors in the relative scalings between different locations, and even cells that have been omitted or falsely detected. Further, the behavior of the EMD with respect to the noise level confirms what we observed in Fig. 5, in which the progressive increase of the noise level seems to have no effect up to a breaking point. Figs. 5 and 6, then, appear to suggest that this breaking point is more related to the inverse problem at hand than to any metric in particular.

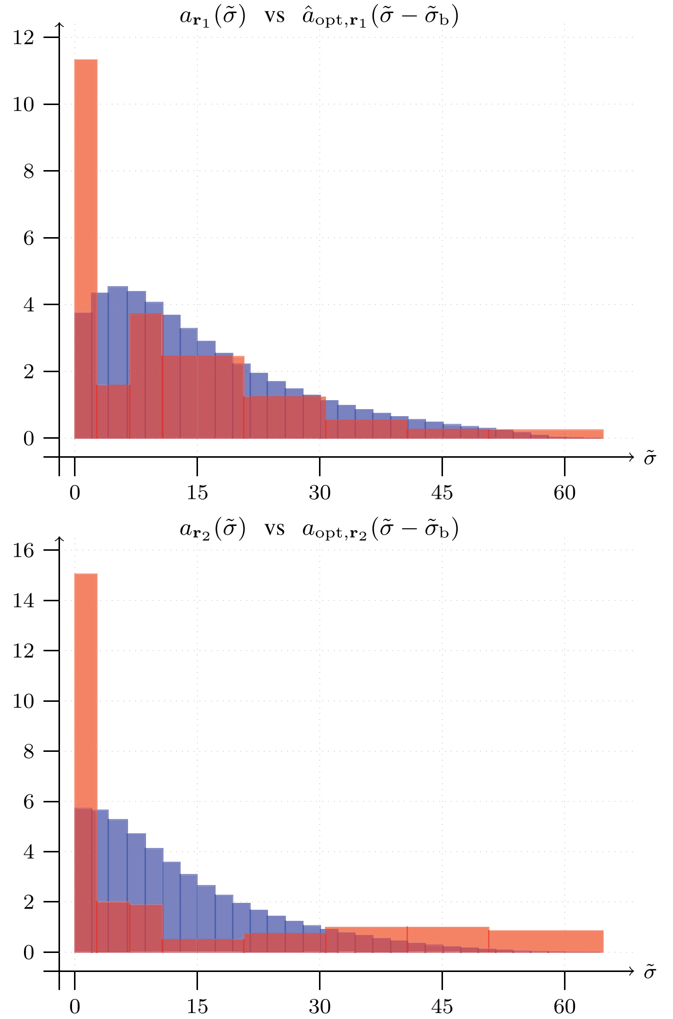


Fig. 7. Two extreme examples of the recovery of $a_{\mathbf{r}}(\tilde{\sigma})$ in simulated spots in different simulation conditions. In blue, $a_{\mathbf{r}}(\tilde{\sigma})$ that is used to simulate the particular spot, with generation parameters as in Table II. In red, $\hat{a}_{\mathrm{opt},\mathbf{r}}(\tilde{\sigma})$ that is recovered by Algorithm 3 with the parameters in Table III, $\lambda = 0.5$, and using the kernel approximations g_k^{br1}. Above, recovery for a cell in an image with $N_{\mathrm{c}} = 1$ and noise level 1. Below, recovery for a well-detected cell in an image with $N_{\mathrm{c}} = 1250$ and noise level 4. The two profiles were normalized to integrate to the same total secretion.

3) Recovery of the Third Dimension: To finalize this section, we transcend the purpose of localization and present in Fig. 7 two examples of the recovery of the PSDR's behavior over $\tilde{\sigma}$, i.e., $a_{\mathbf{r}}(\tilde{\sigma})$ for some location $\mathbf{r} \in \mathbb{R}^2$. There, we can see that the quality of this recovery depends highly on the simulation conditions. On one hand, in complete absence of interference, i.e., in an image with $N_{\mathrm{c}} = 1$, and with noise level 1, this recovery partly captures some of the traits of the real curve. In particular, although it exhibits important errors for lower

$\tilde{\sigma}$ s, it correctly captures the decay of the amount of secretion from $k = 3$ onwards. On the other hand, with many interfering sources ($N_{\rm c} = 1250$) and noise level 4, the information on the $\tilde{\sigma}$-dependence of the PSDR is completely lost. Although we will not explore this any further in this paper, note that the positive weighting function ξ in (1) introduced in Part I [19] could be used to correct systematic errors in the estimation of the PSDR and its profile over $\tilde{\sigma}$, like the apparent systematic overestimation of the first bin in Fig. 7.

IV. Discussion

Throughout this two-part paper, we have: 1) proposed an observation model in function spaces for image measurements of a 3D reaction-diffusion-adsorption-desorption system; 2) provided results that fully characterize our observation model with respect to physical parameters and allow the generation of synthetic data; 3) proposed an optimization problem in function spaces for inverse diffusion when the reactive term is spatially localized and temporally continuous; 4) proposed sound methodology for solving the aforementioned optimization problem; 5) contributed a novel proximal operator of the sum of two functions, i.e., that of the non-negative group-sparsity regularizer; 6) provided a discretization scheme that leads to practical, easy-to-approximate algorithms for synthesis and analysis of data based on our methods and; 7) thoroughly examined the results of our optimization algorithm in terms of operational performance metrics and distributional recovery metrics.

The proposed algorithm provides high-performing, unmatched SL detection results across a wide range of realistic experimental conditions with remarkable and unique robustness to additive noise. Moreover, the source location estimates are very accurate, even when the observed spot is the result of the combined emissions of several close sources. Additionally, although our algorithm requires an hyper-parameter λ, it is very robust to its choice, which makes it a good candidate for practical use.

Our study is not without limitations. In particular, because much of what concerns discretization of ill-posed functional inverse problems is yet unknown, we provide no guideline for discretizations of our functional methods different than our own. In fact, even our own discretization is argued intuitively, and theoretical results to strongly support it are left for further research. Furthermore, because we focus on the analysis of the proposed optimization framework, we disregard the discussion of convergence, relying only on the theoretical results on the convergence rate. In practice, however, theoretically sound approaches to speeding up the convergence of proximal-gradient-based algorithms are available [43], [44], and many heuristics can substantially reduce computations without a substantial loss in SL detection performance. In particular, the masking function $\mu(\mathbf{r})$ in our optimization framework can be modified adaptively after some iterations in order to discard regions that appear to have no cells, and thereby focus the algorithmic effort on more promising areas. Finally, because the computations involved in generating synthetic data are substantial, we limit the resolution of the underlying source locations to 1 pix, and obtain the discretized kernels for generation from the hypothesis that cells are pixel-centered. This could bias our analysis, since kernels within our algorithms are computed analogously. However, we consider that this is unlikely, because the different discretization of the $\tilde{\sigma}$-dimension in synthesis and analysis affects the kernels greatly, and because experimentation on real data yields similarly impressive results.

In our work, we have also encountered paths for future research efforts. While discussing discretizations, we have suggested representations with either thinner spatial grids or off-the-grid solutions to obtain super-resolution location accuracy. Throughout the paper, we have suggested that further improvements in the estimation of the PSDR could enable the study and assessment of the per-cell secretion in a biochemical assay. These improvements could plausibly be achieved through the weighting function $\xi(\sigma)$ that controls the group-sparsity regularizer, which is supported in all of our theoretical results. Finally, through our discretized algorithm in this paper, we have provided empirical evidence that tensor-based modeling of a matrix observation, through adequate group-sparsity coupling and non-negativity constraints, is a viable option for the reconstruction of highly-structured images.

Appendix

Constraints and Regularization, Proximal Operators

In this appendix, we will provide and prove the results on proximal operators upon which the proposed regularized algorithm relies. These will relate to the analysis of the functional f in (4), which represents both the regularizer and the constraint imposed in (1). Because this appendix includes the most technical functional-analytic derivations in the paper, we will first introduce some extra notations, and urge the interested reader to explore [24], [37] for details on optimization in function spaces and relevant references.

A. Notation

Consider this section a continuation of Section I-B. For any specific functional $f : \mathcal{Y} \to \mathbb{R}$, $f_- : \mathcal{Y} \to \mathbb{R}$ is its negative part, i.e., $f_-(y) = \min\{f(y), 0\}, \forall y \in \mathcal{Y}$, and we have that $f = f_+ + f_-$.

When discussing a Hilbert space $\mathcal{X}$, $\mathcal{X}^*$ is its dual space, and for any $x^* \in \mathcal{X}^*$, $r_{x^*} \in \mathcal{X}$ is its Riesz representation, i.e. $\langle x^*, x \rangle_{\mathcal{X}} = (r_{x^*}|x)_{\mathcal{X}}, \forall x \in \mathcal{X}$. Further, $x^*_{\rm p} \in \mathcal{X}^*$ (and $x^*_{\rm n} \in \mathcal{X}^*$) are the linear continuous functionals represented by $[r_{x^*}]_+$ (and $[r_{x^*}]_-$, respectively), and $x^* = x^*_{\rm p} + x^*_{\rm n}$. We refer to $x^*_{\rm p}$ and $x^*_{\rm n}$ as dual-positive and dual-negative parts, respectively.

When discussing a normed functional space, and given any strictly positive weight function, i.e., $\xi \in \mathcal{X}_+$ such that $1/\xi = \xi^{-1} \in \mathcal{X}_+$, and $\gamma > 0$, $\bar{\mathcal{B}}_\xi(\gamma) = \left\{x \in \mathcal{X} : \left\|\xi^{-1}x\right\|_{\mathcal{X}} \leq \gamma\right\}$ is the closed ellipsoid with constant ξ^{-1}-weighted norm under γ and $\bar{\mathcal{B}}^*_\xi(\gamma) = \left\{x^* \in \mathcal{X}^* : \left\|\xi^{-1}r_{x^*}\right\|_{\mathcal{X}} \leq \gamma\right\}$ is the closed dual ellipsoid with ξ^{-1}-weighted dual norm under γ. Additionally, $\bar{\mathcal{B}}(\gamma)$ is the closed ball in $\mathcal{X}$ with norm under γ. Finally, for any convex set $\mathcal{Z} \subset \mathcal{X}$, $\mathrm{P}_{\mathcal{Z}} : \mathcal{X} \to \mathcal{Z}$ is the projection operator onto it, i.e.,

$$\mathrm{P}_{\mathcal{Z}}\left[x\right] = \arg\min_{y\in\mathcal{Z}} \left[\left\|y - x\right\|_{\mathcal{X}}^2\right] .$$

B. Proximal Operator of the Positively-Constrained Weighted Norm in $L^2\ (\aleph)$

Throughout this section, recall the weighting function $\xi \in \mathrm{L}_+^\infty\left[0, \sigma_{\max}\right]$ introduced in Section I and [19, Section III], and recall that we use $\aleph = \mathrm{supp}\left(\xi\right)$ and $\aleph^{\mathsf{c}} = [0, \sigma_{\max}]\backslash\aleph$. Additionally, let $\mathcal{X} = \mathrm{L}^2\left(\aleph\right)$.

In Definition 1, we introduce the functional that characterizes the behavior of the constraint set and the regularizer in (1) in the σ-dimension.

Definition 1 (Non-negative weighted norm in $\mathcal{X}$): Define the the functional

$$\begin{aligned} \vartheta : \mathcal{X} &\to \bar{\mathbb{R}}_+ && \text{(8a)}\\ x &\mapsto \left\|\xi x\right\|_{\mathcal{X}} + \delta_{\mathcal{X}_+}\left(x\right), \forall x \in \mathcal{X}. && \text{(8b)} \end{aligned}$$

The main results of this appendix, which will be presented in Lemmas 3 and 4, will provide the value of the proximal operator of $\gamma\vartheta$, i.e., $\mathrm{prox}_{\gamma\vartheta}\left(x\right), \forall x \in \mathcal{X}, \forall\gamma > 0$. This results are used in the proofs of Theorems 1 and 2, which are presented at the end of the appendix. In order to derive Lemmas 3 and 4, we will follow a path similar to the classical proof of the proximal operator of a norm. We will first find the convex conjugate functional $\left(\gamma\vartheta\right)^*$ in Lemma 1. Then, we will derive its proximal operator $\mathrm{prox}_{\left(\gamma\vartheta\right)^*}(x^*), \forall x^* \in \mathcal{X}^*, \forall\gamma > 0$ in Lemma 2. Finally, we will use this result and Moreau's identity to lead us into Lemmas 3 and 4.

Lemma 1 (Fenchel conjugate of the scaled, non-negative weighted norm in $\mathcal{X}$): Consider the functional ϑ in Definition 1. Then, $\forall\gamma > 0$, we have that the convex conjugate functional of $\gamma\vartheta$ is

$$\begin{aligned} \left(\gamma\vartheta\right)^* : \mathcal{X}^* &\to \bar{\mathbb{R}}\\ x^* &\mapsto \delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*)\,, \forall x^* \in \mathcal{X}^*, \end{aligned}$$

with $\bar{\mathcal{B}}_\xi^*\left(\gamma\right)$ as defined in the previous section.

Proof: Here, we will instead show that the Fenchel conjugate of $\delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*)$ is the functional $\hat{\vartheta} : \mathcal{X} \to \bar{\mathbb{R}}$ such that

$$\hat{\vartheta} = \left[\delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*)\right]^* = \gamma\vartheta\,.$$

The Fenchel-Moreau theorem then allows us to conclude that, because ϑ is convex, proper and lower semi-continuous, $\delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*)$ is the Fenchel conjugate of $\gamma\vartheta$.

Starting now from the definition of $\hat{\vartheta}$ we obtain

$$\begin{aligned} \hat{\vartheta}(x) &= \sup_{x^*\in\mathcal{X}^*}\left\{\langle x^*, x\rangle_{\mathcal{X}} - \delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*)\right\}\\ &= \sup_{x^*\in\mathcal{X}^*}\left\{\langle x_{\mathrm{n}}^*, x\rangle_{\mathcal{X}} + \left\langle x_{\mathrm{p}}^*, x\right\rangle_{\mathcal{X}} - \delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*)\right\} . \end{aligned}$$

Here, we can readily determine that, if $x \notin \mathcal{X}_+$, $\hat{\vartheta}(x) = +\infty$. Indeed, if $\exists S \subset \aleph$ such that $x < 0$ a.e. in S, we have that for $\hat{\vartheta}(x) \geq \sup_{x^*\in\mathcal{X}^*}\langle x_{\mathrm{n}}^*, x\rangle_{\mathcal{X}} \geq \sup_{K<0} K\int_S x = +\infty$.

We continue by noting that, if $x \in \mathcal{X}_+$, then $\langle x_{\mathrm{n}}^*, x\rangle_{\mathcal{X}} \leq 0$, and thus, it will be enough to consider the case $x^* = x_{\mathrm{p}}^*$, i.e., $x_{\mathrm{n}}^* = 0$, to determine $\hat{\vartheta}(x)$. Therefore, for any $x \in \mathcal{X}_+$ we can use the Cauchy-Schwartz inequality to show that

$$\begin{aligned} \hat{\vartheta}(x) &= \sup_{x^*\in\mathcal{X}^*}\left\{\left\langle x_{\mathrm{p}}^*, x\right\rangle_{\mathcal{X}} - \delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*)\right\}\\ &= \sup_{x^*\in\mathcal{X}^*}\left\{\langle x^*, x\rangle_{\mathcal{X}} - \delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x^*)\right\}\\ &= \sup_{x^*\in\bar{\mathcal{B}}_\xi^*(\gamma)}\left\{\langle x^*, x\rangle_{\mathcal{X}}\right\}\\ &= \sup_{r_{x^*}\in\bar{\mathcal{B}}_\xi(\gamma)}\left\{\left(\xi^{-1}r_{x^*}|\xi x\right)_{\mathcal{X}}\right\} = \gamma\left\|\xi x\right\|_{\mathcal{X}}\ . \end{aligned}$$

In conclusion, then,

$$\begin{aligned} \hat{\vartheta}(x) &= \left\{\begin{matrix} +\infty & \text{if } x \notin \mathcal{X}_+\,, \\ \gamma\left\|\xi x\right\|_{\mathcal{X}} & \text{if } x \in \mathcal{X}. \end{matrix}\right\}\\ &= \gamma\left\|\xi x\right\|_{\mathcal{X}} + \delta_{\mathcal{X}_+}\left(x\right) = \gamma\vartheta\,, \end{aligned}$$

which finishes our proof. ■

Similarly to what happens with the dual of a norm, the dual functional $\left(\gamma\vartheta\right)^*$ is a simple indicator. This makes its proximal operator in Lemma 2 a combination of simple, standard operations, such as dual-positive and dual-negative parts, and projections onto convex sets.

Lemma 2 (Projection of the positive part on the dual ellipsoid): Consider the functional

$$\begin{aligned} \zeta : \mathcal{X}^* &\to \{0, +\infty\}\\ x^* &\mapsto \delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(x_{\mathrm{p}}^*), \forall x^* \in \mathcal{X}^*\,, \end{aligned}$$

i.e., $\zeta = (\gamma\vartheta)^*$. Then,

$$\mathrm{prox}_\zeta(x^*) = x_{\mathrm{n}}^* + \mathrm{P}_{\bar{\mathcal{B}}_\xi^*(\gamma)}\left[x_{\mathrm{p}}^*\right]\ .$$

Proof: Recall here that obtaining $\mathrm{prox}_\zeta(x^*)$ is obtaining the minimizer of

$$\min_{y^*\in\mathcal{X}^*}\left[\frac{1}{2}\left\|y^* - x^*\right\|_{\mathcal{X}^*}^2 + \delta_{\bar{\mathcal{B}}_\xi^*(\gamma)}(y_{\mathrm{p}}^*)\right]\ . \qquad (9)$$

In this proof, let us refer to this minimizer as y_{opt}^*. The solution to (9) is intuitively simple. On one hand, because the value $\zeta(y^*)$ does not vary with changes in the negative part of y^*, we will have that the minimization of the term $\left\|y^* - x^*\right\|_{\mathcal{X}^*}^2$ will dominate the negative part of the optimal solution and $y_{\mathrm{opt,n}}^* = x_{\mathrm{n}}^*$. On the other hand, the positive part of y^* is only constrained to be in the ellipsoid $\bar{\mathcal{B}}_\xi^*\left(\gamma\right)$ and, thus, the positive part of the solution will be the point at minimum distance from x_{p}^* inside the ellipsoid, i.e., $y_{\mathrm{opt,p}}^* = \mathrm{P}_{\bar{\mathcal{B}}_\xi^*(\gamma)}\left[x_{\mathrm{p}}^*\right]$.

Let us now formalize this by considering any element $x^* \in \mathcal{X}^*$, and letting $N_{x^*} = \{\sigma \in \aleph : r_{x^*}(\sigma) < 0\} = \mathrm{supp}\left(r_{x_n^*}\right)$. For any $y^* \in \mathcal{X}^*$, let $p_{y^*}, n_{y^*} \in \mathcal{X}$ be such that

$$\mathrm{supp}\left(p_{y^*}\right) \subset N_{x^*}^{\mathsf{c}},\ \mathrm{supp}\left(n_{y^*}\right) \subset N_{x^*}, \qquad (10)$$

with $N_{x^*}^{\mathsf{c}} = \aleph \setminus N_{x^*}$, and

$$r_{y^*} = p_{y^*} + n_{y^*}\,. \qquad (11)$$

Then,

$$\begin{aligned}\|y^* - x^*\|^2_{\mathcal{X}^*} &= \int_{N^{\mathsf{c}}_{x^*}} \left(r_{x^*_{\mathrm{p}}} - p_{y^*}\right)^2 \\ &\quad + \int_{N_{x^*}} \left(r_{x^*_{\mathrm{n}}} - n_{y^*}\right)^2 \\ &= \left\|r_{x^*_{\mathrm{p}}} - p_{y^*}\right\|^2_{\mathcal{X}} + \left\|r_{x^*_{\mathrm{n}}} - n_{y^*}\right\|^2_{\mathcal{X}} \,.\end{aligned}$$

Therefore, (9) is equivalent to

$$\min_{\substack{p_{y^*}, n_{y^*} \in \mathcal{X} \\ \text{s.t.} \left[p_{y^*}\right]_+ + \left[n_{y^*}\right]_+ \in \bar{\mathcal{B}}_\xi(\gamma)}} \left[\frac{1}{2}\left\|r_{x^*_{\mathrm{p}}} - p_{y^*}\right\|^2_{\mathcal{X}} + \frac{1}{2}\left\|r_{x^*_{\mathrm{n}}} - n_{y^*}\right\|^2_{\mathcal{X}}\right]$$

as long as (10) and (11) are fulfilled.

We will now prove that $n_{y^*_{\mathrm{opt}}} \le 0$ a.e. in $\aleph$, which will decouple the minimization of the two summands in the problem above. Assume that $y^*_{\mathrm{opt}} \in \mathcal{X}^*$ is an optimal point of (9) that does not fulfill this condition, i.e., that if

$$\Theta = \left\{\sigma \in N_{x^*} : n_{y^*_{\mathrm{opt}}}(\sigma) > 0\right\}, \text{ then } \int_\Theta n_{y^*_{\mathrm{opt}}} > 0 \,.$$

Let $y^*_1 \in \mathcal{X}^*$ such that $p_{y^*_1} = p_{y^*_{\mathrm{opt}}}$, and $n_{y^*_1} = \left[n_{y^*_{\mathrm{opt}}}\right]_-$. Then, because y^*_{opt} was a feasible point, i.e., $y^*_{\mathrm{opt,p}} \in \bar{\mathcal{B}}^*_\xi(\gamma)$, we have that

$$\begin{aligned}\gamma^2 &\ge \int_\aleph \xi^{-2}\left(\left[p_{y^*_{\mathrm{opt}}}\right]_+ + \left[n_{y^*_{\mathrm{opt}}}\right]_+\right)^2 \\ &\ge \int_\aleph \xi^{-2}\left[p_{y^*_{\mathrm{opt}}}\right]^2_+ = \|y^*_{1,\mathrm{p}}\|^2_{\mathcal{X}^*} \,,\end{aligned}$$

i.e., $y^*_{1,\mathrm{p}} \in \bar{\mathcal{B}}^*_\xi(\gamma)$ and y^*_1 is a feasible point. Moreover, $\forall \sigma \in \Theta$, we have that $\left|r_{x^*_{\mathrm{n}}}(\sigma) - n_{y^*_{\mathrm{opt}}}(\sigma)\right| > \left|r_{x^*_{\mathrm{n}}}(\sigma)\right|$ and thus

$$\begin{aligned}\left\|r_{x^*_{\mathrm{n}}} - n_{y^*_{\mathrm{opt}}}\right\|^2_{\mathcal{X}} &= \int_{N_{x^*}} \left(r_{x^*_{\mathrm{n}}} - n_{y^*_{\mathrm{opt}}}\right)^2 \\ &> \int_{N_{x^*}\setminus\Theta} \left(r_{x^*_{\mathrm{n}}} - n_{y^*_{\mathrm{opt}}}\right)^2 + \int_\Theta r^2_{x^*_{\mathrm{n}}} \\ &= \left\|r_{x^*_{\mathrm{n}}} - n_{y^*_1}\right\|^2_{\mathcal{X}} \,, \end{aligned} \tag{12}$$

which implies that $\left\|y^*_{\mathrm{opt}} - x^*\right\|^2_{\mathcal{X}^*} > \|y^*_1 - x^*\|^2_{\mathcal{X}^*}$. This contradicts the optimality of y^*_{opt}. Thus, an optimal point y^*_{opt} must fulfill $n_{y^*_{\mathrm{opt}}} \le 0$ a.e. in $\aleph$.

Therefore, (13) is equivalent to (9), as long as conditions (10) and (11) are fulfilled.

$$\min_{\substack{p_{y^*} \in \mathcal{X} \\ \text{s.t.} \left[p_{y^*}\right]_+ \in \bar{\mathcal{B}}_\xi(\gamma)}} \left[\frac{1}{2}\left\|r_{x^*_{\mathrm{p}}} - p_{y^*}\right\|^2_{\mathcal{X}}\right] \tag{13a}$$

$$\min_{n_{y^*} \in \mathcal{X}} \left[\frac{1}{2}\left\|r_{x^*_{\mathrm{n}}} - n_{y^*}\right\|^2_{\mathcal{X}}\right] \tag{13b}$$

(13b) is an unconstrained norm minimization, and has its minimum at $n_{y^*_{\mathrm{opt}}} = r_{x^*_{\mathrm{n}}}$, which fulfills (10). Using an argument parallel to the one that lead to (12), we have that $r_{x^*_{\mathrm{p}}} \ge 0$ a.e. in $\aleph$ implies that $p_{y^*_{\mathrm{opt}}} \ge 0$ a.e. in $\aleph$ too. Thus, (9) is equivalent to

$$\min_{p_{y^*} \in \bar{\mathcal{B}}_\xi(\gamma)} \left[\frac{1}{2}\left\|r_{x^*_{\mathrm{p}}} - p_{y^*}\right\|^2_{\mathcal{X}}\right] \,,$$

and thus, $p_{y^*_{\mathrm{opt}}} = \mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}[r_{x^*_{\mathrm{p}}}]$. In Property 2, we obtain an expression for $\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}[x]$ for any $x \in \mathcal{X}$ that shows that $\mathrm{supp}(\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}[x]) \subset \mathrm{supp}(x)$ and, thus, (10) is fulfilled. Then, the solution to (9) is given by (11) as the $y^*_{\mathrm{opt}} \in \mathcal{X}^*$ represented by $r_{y^*_{\mathrm{opt}}} = r_{x^*_{\mathrm{n}}} + \mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}[r_{x^*_{\mathrm{p}}}]$, i.e.,

$$y^*_{\mathrm{opt}} = x^*_{\mathrm{n}} + \mathrm{P}_{\bar{\mathcal{B}}^*_\xi(\gamma)}\left[x^*_{\mathrm{p}}\right] \,. \qquad \blacksquare$$

We now can use the relation between the proximal operator of a functional and that of its convex conjugate to finally achieve the desired result in Lemma 3.

Lemma 3 (Proximal operator of the scaled, non-negative weighted norm in $\mathcal{X}$): Consider the functional ϑ in Definition 1. Then, $\forall \gamma > 0$, we have that the proximal operator of the functional $\gamma\vartheta$ is

$$\mathrm{prox}_{\gamma\vartheta}(x) = x_+ - \mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}\left[x_+\right], \forall x \in \mathcal{X} \,.$$

Proof: Lemmas 1 and 2 grant that $\mathrm{prox}_{(\gamma\vartheta)^*}(x^*) = x^*_{\mathrm{n}} + \mathrm{P}_{\bar{\mathcal{B}}^*_\xi(\gamma)}\left[x^*_{\mathrm{p}}\right]$. A well-known generalization of Moreau's decomposition theorem for projection on convex cones in Hilbert spaces is that

$$\mathrm{prox}_{\gamma\vartheta}(x) + \mathrm{prox}_{(\gamma\vartheta)^*}(x) = x \,. \tag{14}$$

Note here that we abuse the notation by identifying $\mathcal{X}$ with its dual $\mathcal{X}^*$ and $\mathrm{prox}_{(\gamma\vartheta)^*}(x)$ with $r_{\mathrm{prox}_{(\gamma\vartheta)^*}(x^*)} \in \mathcal{X}$ such that $x^* \in \mathcal{X}^*$ is represented by $r_{x^*} = x$. Directly from (14), then, we obtain that

$$\begin{aligned}\mathrm{prox}_{\gamma\vartheta}(x) &= x - \mathrm{prox}_{(\gamma\vartheta)^*}(x) \\ &= x_+ - \mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}\left[x_+\right] . \qquad \blacksquare\end{aligned}$$

Although we now have our result compactly expressed in terms of simple, known operations, the inherent optimization problem in the term $\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}\left[x_+\right]$ is known to have no closed-form solution. For completeness, we include this result in Property 2.

Property 2 (Projection on an ellipsoid): The projection of a functional $x \in \mathcal{X}$ onto the closed ellipsoid $\bar{\mathcal{B}}_\xi(\gamma)$ is

$$\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}[x] = \begin{cases} x & \text{if } x \in \bar{\mathcal{B}}_\xi(\gamma) \,, \\ \frac{\xi^2}{\xi^2 + 2\lambda} x & \text{if } x \in \mathcal{X} \setminus \bar{\mathcal{B}}_\xi(\gamma) \,, \end{cases}$$

with $\lambda \ge 0$ such that

$$\left\|\frac{\xi}{\xi^2 + 2\lambda} x\right\|_{\mathcal{X}} = \gamma \,.$$

Proof: Recall here that the projection operator is defined as

$$\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma)}[x] = \arg\min_{y \in \bar{\mathcal{B}}_\xi(\gamma)} \left[\frac{1}{2}\|x - y\|^2_{\mathcal{X}}\right] . \tag{15}$$

Because $\mathcal{X}$ is complete and $\bar{\mathcal{B}}_\xi(\gamma)$ is convex and closed, the projection operator is well defined and strong Lagrange duality is granted. Note that the convexity of $\bar{\mathcal{B}}_\xi(\gamma)$ is granted by the convexity of the weighted norm $\left\|\xi^{-1}\cdot\right\|_{\mathcal{X}}$, which follows directly from the convexity of the norm $\|\cdot\|_{\mathcal{X}}$.

The Lagrangian for this problem is

$$\begin{aligned} L(y,\lambda) &= \frac{1}{2}\left\|x-y\right\|_{\mathcal{X}}^2+\lambda\left(\left\|\xi^{-1}y\right\|_{\mathcal{X}}^2-\gamma^2\right) \\ &= \left(y\left[\frac{1}{2}+\lambda\xi^{-2}\right]|y\right)_{\mathcal{X}}+\frac{1}{2}\left(x|x\right)_{\mathcal{X}}-\left(x|y\right)_{\mathcal{X}}-\lambda\gamma^2 \end{aligned}$$

with $\lambda\geq 0$. Because the Lagrangian $L(y,\lambda)$ is convex and Frchet differentiable with respect to $y\in\mathcal{X}$, and its Frchet derivative is $\nabla_y L(y,\lambda)=2y\left[\frac{1}{2}+\lambda\xi^{-2}\right]-x$, its minimizer $y_{\mathrm{opt}}\in\mathcal{X}$ is

$$y_{\mathrm{opt}}(\lambda)=\frac{1}{2}\frac{x}{\frac{1}{2}+\lambda\xi^{-2}}=\frac{\xi^2}{\xi^2+2\lambda}x\,.$$

The dual function for (15) is

$$\begin{aligned} h(\lambda) =&\ L(y_{\mathrm{opt}},\lambda) \\ =&\ -\lambda\gamma^2+\frac{1}{2}\left[\left\|x\right\|_{\mathcal{X}}^2-\left(x|\frac{1}{2}\frac{x}{\frac{1}{2}+\lambda\xi^{-2}}\right)_{\mathcal{X}}\right] \\ =&\ \frac{\left\|x\right\|_{\mathcal{X}}^2}{2}-\lambda\gamma^2-\int_{\aleph}\frac{x^2}{4}\frac{1}{\frac{1}{2}+\lambda\xi^{-2}}\,, \end{aligned}$$

which is concave in $\lambda\geq 0$ and, thus, has its maximum at either $\lambda_{\mathrm{opt},1}=0$ or at that $\lambda_{\mathrm{opt},2}$ that yields

$$\frac{\partial}{\partial\lambda}h=-\gamma^2+\int_{\aleph}\frac{x^2}{4}\frac{\xi^{-2}}{\left(\frac{1}{2}+\lambda_{\mathrm{opt},2}\xi^{-2}\right)^2}=0\,,$$

i.e., $\left\|\xi^{-1}y_{\mathrm{opt}}(\lambda_{\mathrm{opt},2})\right\|_{\mathcal{X}}=\gamma$. Generally, the value of $\lambda_{\mathrm{opt},2}$ cannot be obtained in closed form.

If $x\in\bar{\mathcal{B}}_{\xi}(\gamma)$, we know that the optimal value for (15) is 0 and is achieved at $y_{\mathrm{opt}}=x$, which implies that the optimal Lagrange multiplier is $\lambda=\lambda_{\mathrm{opt},1}=0$. If $x\in\mathcal{X}\setminus\bar{\mathcal{B}}_{\xi}(\gamma)$, we know that the optimal value for (15) must be larger than zero, which by strong duality implies that $\lambda\neq 0$ and, thus, that the optimal Lagrange multiplier is $\lambda=\lambda_{\mathrm{opt},2}$ and the optimal primal point is $y_{\mathrm{opt}}(\lambda_{\mathrm{opt},2})$, which is primal-feasible by definition. ■

This result determines the shape of $\mathrm{P}_{\bar{\mathcal{B}}_{\xi}(\gamma)}\left[x_+\right]$, but it does not give a closed-form expression for it. To find this projection, the value of λ in Property 2 has to be found. Several numerical methods have been developed to find this value or otherwise compute the projection on an ellipsoid [45]. Using these in the context of our problem, however, is outside the scope of our paper. We opt instead for particularizing in Property 3 to cases in which the weighting function is constant a.e. in $\aleph$. This makes the projection to be computed, without loss of generality, $\mathrm{P}_{\bar{\mathcal{B}}(\gamma)}\left[x_+\right]$, the projection onto a closed ball in $\mathcal{X}$.

Property 3 (Projection on a ball): The projection of a functional $x\in\mathcal{X}$ onto the closed ball of norm under γ, i.e., $\bar{\mathcal{B}}(\gamma)$, is

$$\mathrm{P}_{\bar{\mathcal{B}}(\gamma)}\left[x\right]=\begin{cases} x & \text{if } x\in\bar{\mathcal{B}}(\gamma)\,, \\ \frac{\gamma}{\|x\|_{\mathcal{X}}}x & \text{if } x\in\bar{\mathcal{B}}(\gamma)^{\mathsf{c}}\,. \end{cases}$$

Proof: Note that this is nothing but a particular case of Property 2 in which $\xi=1$ a.e. in $\aleph$. Then, the case $x\in\bar{\mathcal{B}}(\gamma)$ is trivial. For $x\in\bar{\mathcal{B}}(\gamma)^{\mathsf{c}}$, the equation for $\lambda\geq 0$ in Property 2 can be solved in closed form, yielding

$$\left\|\frac{x}{1+2\lambda}\right\|_{\mathcal{X}}=\left|\frac{1}{1+2\lambda}\right|\|x\|_{\mathcal{X}}=\frac{1}{1+2\lambda}\|x\|_{\mathcal{X}}=\gamma\,,$$

i.e., $\lambda=\frac{1}{2}\left(\frac{\|x\|_{\mathcal{X}}}{\gamma}-1\right)$, and

$$y_{\mathrm{opt}}(\lambda)=\frac{\xi^2 x}{\xi^2+2\lambda}=\frac{x}{1+2\frac{\frac{\|x\|_{\mathcal{X}}}{\gamma}-1}{2}}=\frac{\gamma}{\|x\|_{\mathcal{X}}}x.$$ ■

This allows us to obtain a closed-form version of Lemma 3 for this specific choice of ξ, i.e., Lemma 4.

Lemma 4 (Proximal operator of the scaled, non-negative norm in $\mathcal{X}$): Consider the functional ϑ in Definition 1 when $\xi=1$ a.e. in $\aleph$. Then, $\forall\gamma>0$, we have that the proximal operator of the functional $\gamma\vartheta$ is

$$\mathrm{prox}_{\gamma\vartheta}(x)=x_+\left(1-\frac{\gamma}{\|x_+\|_{\mathcal{X}}}\right)_+\,.$$

Proof: Using Lemma 3 and Property 3 we obtain that

$$\begin{aligned} \mathrm{prox}_{\gamma\vartheta}(x) =&\ x_+-\mathrm{P}_{\bar{\mathcal{B}}(\gamma)}\left[x_+\right] \\ =&\ \begin{cases} 0 & \text{if } \|x_+\|_{\mathcal{X}}/\gamma\leq 1, \\ x_+\left(1-\frac{\gamma}{\|x_+\|_{\mathcal{X}}}\right) & \text{if } \|x_+\|_{\mathcal{X}}/\gamma>1, \end{cases} \\ =&\ x_+\left(1-\frac{\gamma}{\|x_+\|_{\mathcal{X}}}\right)_+\,. \end{aligned}$$ ■

We now use the results above to prove Theorems 1 and 2, which constitute the backbone of Algorithm 2.

Proof - Theorem 1 (Proximal operator of the non-negative weighted group-sparsity regularizer): Consider the functional ϑ in Definition 1. Then, recalling the functional f in (4), we have

$$\begin{aligned} \gamma f(a) &= \delta_{\mathcal{A}_+}(a)+\gamma\lambda\left\|\|\xi a_{\mathbf{r}}\|_{\mathcal{X}}\right\|_{\mathrm{L}^1(\mathbb{R}^2)} && \text{(16a)} \\ &= \delta_{\mathcal{A}_+}(a)+\gamma\lambda\int_{\mathbb{R}^2}\|\xi a_{\aleph,\mathbf{r}}\|_{\mathcal{X}}\,\mathrm{d}\mathbf{r} && \text{(16b)} \\ &= \int_{\mathbb{R}^2}\left(\delta_{\mathrm{L}_+^2[0,\sigma_{\max}]}(a_{\mathbf{r}})+\gamma\lambda\|\xi a_{\aleph,\mathbf{r}}\|_{\mathcal{X}}\right)\mathrm{d}\mathbf{r} && \text{(17a)} \\ &= \int_{\mathbb{R}^2}\left(\delta_{\mathrm{L}_+^2(\aleph^{\mathsf{c}})}(a_{\aleph^{\mathsf{c}},\mathbf{r}})+\gamma\lambda\vartheta\left(a_{\aleph,\mathbf{r}}\right)\right)\mathrm{d}\mathbf{r}\,. && \text{(17b)} \end{aligned}$$

Here, (16) uses that $\xi a_{\mathbf{r}}=0$ a.e. in $\aleph^{\mathsf{c}}$, and (17) uses that $a\in\mathcal{A}_+$ is equivalent to $a_{\mathbf{r}}\in\mathrm{L}_+^2[0,\sigma_{\max}]$ for almost any $\mathbf{r}\in\mathbb{R}^2$, and that is equivalent to $a_{\aleph^{\mathsf{c}},\mathbf{r}}\in\mathrm{L}_+^2(\aleph^{\mathsf{c}})$ and $a_{\aleph,\mathbf{r}}\in\mathrm{L}_+^2(\aleph)=\mathcal{X}_+$ for almost any $\mathbf{r}\in\mathbb{R}^2$.

$\mathrm{prox}_{\gamma f}(a)$ is the minimizer to

$$\min_{b\in\mathcal{A}}\left[\frac{1}{2}\|b-a\|_{\mathcal{A}}^2+\gamma f(b)\right]\,,$$

or, equivalently,

$$\min_{b\in\mathcal{A}}\left[\int_{\mathbb{R}^2}\left(\frac{1}{2}\left\|b_{\mathbf{r},\aleph^{\mathsf{c}}}-a_{\aleph^{\mathsf{c}},\mathbf{r}}\right\|^2_{\mathrm{L}^2(\aleph^{\mathsf{c}})}+\delta_{\mathrm{L}^2_+(\aleph^{\mathsf{c}})}(b_{\aleph^{\mathsf{c}},\mathbf{r}})\right.\right.$$
$$\left.\left.+\frac{1}{2}\left\|b_{\mathbf{r},\aleph}-a_{\aleph,\mathbf{r}}\right\|^2_{\mathcal{X}}+\gamma\lambda\vartheta\left(b_{\aleph,\mathbf{r}}\right)\right)\mathrm{d}\mathbf{r}\right].$$

By linearity of the integral, then, the optimization of b for $\sigma\in\aleph$ and $\sigma\in\aleph^{\mathsf{c}}$ is completely decoupled. Furthermore, if obtaining the minimizer $b_{\mathbf{r},\mathrm{opt}}\in\mathrm{L}[0,\sigma_{\max}]$ of the term inside the integral for each $\mathbf{r}\in\mathbb{R}^2$ and constructing b_{opt} such that $b_{\mathrm{opt}}(\sigma,\mathbf{r})=b_{\mathbf{r},\mathrm{opt}}(\sigma),\forall\sigma\in[0,\sigma_{\max}]$ yields $b_{\mathrm{opt}}\in\mathcal{A}$, b_{opt} will be optimal with respect to the problem above. In this light, we consider first the optimization for $\sigma\in\aleph^{\mathsf{c}}$ and a specific $\mathbf{r}\in\mathbb{R}^2$, i.e.,

$$\arg\min_{b_{\mathbf{r},\aleph^{\mathsf{c}}}\in\mathrm{L}^2_+(\aleph^{\mathsf{c}})}\left[\frac{1}{2}\left\|b_{\mathbf{r},\aleph^{\mathsf{c}}}-a_{\aleph^{\mathsf{c}},\mathbf{r}}\right\|^2_{\mathrm{L}^2(\aleph^{\mathsf{c}})}\right]=\left[a_{\aleph^{\mathsf{c}},\mathbf{r}}\right]_+,$$

and see that it is resolved by a simple non-negative projection. Then, we observe that the optimization for $\sigma\in\aleph$ and a specific $\mathbf{r}\in\mathbb{R}^2$ is of the form considered in Lemma 3, which resolved it to $\left[a_{\aleph,\mathbf{r}}\right]_+-\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma\lambda)}\left[\left[a_{\aleph,\mathbf{r}}\right]_+\right]$. We then have that

$$\begin{aligned}b_{\mathbf{r},\mathrm{opt}}&=\left[a_{\aleph^{\mathsf{c}},\mathbf{r}}\right]_++\left[a_{\aleph,\mathbf{r}}\right]_+-\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma\lambda)}\left[\left[a_{\aleph,\mathbf{r}}\right]_+\right]\\&=\left[a_{\mathbf{r}}\right]_+-\mathrm{P}_{\bar{\mathcal{B}}_\xi(\gamma\lambda)}\left[\left[a_{\aleph,\mathbf{r}}\right]_+\right].\end{aligned}$$

Note that

$$\left\|b_{\mathbf{r},\mathrm{opt}}\right\|^2_{\mathrm{L}^2[0,\sigma_{\max}]}\leq\left\|\left[a_{\mathbf{r}}\right]_+\right\|^2_{\mathrm{L}^2[0,\sigma_{\max}]}\leq\left\|a_{\mathbf{r}}\right\|^2_{\mathrm{L}^2[0,\sigma_{\max}]},\tag{18}$$

and, thus, $a\in\mathcal{A}$ implies that $b_{\mathrm{opt}}\in\mathcal{A}$, which completes the proof of Theorem 1. ■

Proof - Theorem 2 (Proximal operator of the non-negative group-sparsity regularizer on $\aleph$): Using the same proof structure as in Theorem 1, but using Lemma 4 for the optimization for $\sigma\in\aleph$ and a specific $\mathbf{r}\in\mathbb{R}^2$, we obtain that

$$b_{\mathbf{r},\mathrm{opt}}=\left[a_{\aleph^{\mathsf{c}},\mathbf{r}}\right]_++\left[a_{\aleph,\mathbf{r}}\right]_+\left(1-\frac{\gamma\lambda}{\left\|\left[a_{\aleph,\mathbf{r}}\right]_+\right\|_{\mathrm{L}^2(\aleph)}}\right)_+,$$

and (18) implies that $b_{\mathrm{opt}}\in\mathcal{A}$, which concludes the proof. ■

Acknowledgment

Prof. R. Bot, through his lecture series *Recent Advances in Numerical Algorithms for Convex Optimization* given at KTH Royal Institute of Technology in May 2016, was of great help in the development of this research. Dr. C. Smedman provided expert labeling of synthetic data. The authors would like to thank Prof. M. A. T. Figueiredo for first pointing them to [28] during the workshop *Generative models, parameter learning and sparsity* at the Isaac Newton Institute for Mathematical Sciences. The excellent team of anonymous reviewers provided feedback that improved the presentation of our results considerably. Of particular relevance were their pointing to optimal transport and the EMD, and their suggestion of simpler proof techniques for Lemma 1.

References

[1] J.-C. Olivo-Marin, "Extraction of spots in biological images using multiscale products," *Pattern Recognit.*, vol. 35, no. 9, pp. 1989–1996, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320301001273

[2] J. Matthes, L. Groll, and H. B. Keller, "Source localization by spatially distributed electronic noses for advection and diffusion," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1711–1719, May 2005.

[3] A. Hamdi, "Identification of point sources in two-dimensional advection-diffusion-reaction equation: Application to pollution sources in a river. Stationary case," *Inverse Problems Sci. Eng.*, vol. 15, no. 8, pp. 855–870, 2007. [Online]. Available: http://dx.doi.org/10.1080/17415970601162198

[4] J. A. Rebhahn *et al.*, "Automated analysis of two- and three-color fluorescent ELISPOT (Fluorospot) assays for cytokine secretion," *Comput. Methods Programs Biomed.*, vol. 92, no. 1, pp. 54–65, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260708001314

[5] K. Pan, A. Kokaram, J. Hillebrand, and M. Ramaswami, "Gaussian mixture models for spots in microscopy using a new split/merge EM algorithm," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 3645–3648.

[6] I. Smal, M. Loog, W. Niessen, and E. Meijering, "Quantitative comparison of spot detection methods in fluorescence microscopy," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 282–301, Feb. 2010.

[7] Y. Kimori, N. Baba, and N. Morone, "Extended morphological processing: A practical method for automatic spot detection of biological markers from microscopic images," *BMC Bioinf.*, vol. 11, no. 373, 2010. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-373

[8] S. Ram, J. J. Rodrguez, and G. Bosco, "Segmentation and detection of fluorescent 3D spots," *Cytometry Part A*, vol. 81A, no. 3, pp. 198–212, 2012. [Online]. Available: http://dx.doi.org/10.1002/cyto.a.22017

[9] A. Hamdi, "Inverse source problem in a 2D linear evolution transport equation: Detection of pollution source," *Inverse Problems Sci. Eng.*, vol. 20, no. 3, pp. 401–421, 2012. [Online]. Available: http://dx.doi.org/10.1080/17415977.2011.637207

[10] J. Zhao, Y. Li, and S. Du, "A 3-D deconvolution based particle detection method for wide-field microscopy image," in *Proc. 8th Int. Symp. Med. Inf. Commun. Technol.*, Apr. 2014, pp. 1–5.

[11] A. Basset, J. Boulanger, J. Salamero, P. Bouthemy, and C. Kervrann, "Adaptive spot detection with optimal scale selection in fluorescence microscopy images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4512–4527, Nov. 2015.

[12] C. Kervrann, C. O. S. Sorzano, S. T. Acton, J. C. Olivo-Marin, and M. Unser, "A guided tour of selected image processing and analysis methods for fluorescence and electron microscopy," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 1, pp. 6–30, Feb. 2016.

[13] K. Ehrenfried and L. Koop, "A comparison of iterative deconvolution algorithms for the mapping of acoustic sources," *AIAA J.*, vol. 45, no. 7, pp. 1584–1595, Jul. 2006. [Online]. Available: http://dx.doi.org/10.2514/6.2006-2711

[14] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "Resolution issues in soundfield imaging: A multiresolution approach to multiple source localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2015, pp. 1–5.

[15] J. L. Starck, E. Pantin, and F. Murtagh, "Deconvolution in astronomy: A review," *Publications Astronomical Soc. Pacific*, vol. 114, no. 800, pp. 1051–1069, Oct. 2002. [Online]. Available: http://stacks.iop.org/1538-3873/114/i=800/a=1051

[16] J.-F. Giovannelli and A. Coulais, "Positive deconvolution for superimposed extended source and point sources," *Astron. Astrophys.*, vol. 439, no. 1, pp. 401–412, 2005. [Online]. Available: http://dx.doi.org/10.1051/0004-6361:20047011

[17] F. Ternat, P. Daripa, and O. Orellana, "On an inverse problem: Recovery of non-smooth solutions to backward heat equation," *Appl. Math. Modelling*, vol. 36, no. 9, pp. 4003–4019, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0307904X11007086

[18] W. Zhang, F. Ma, and Y. Sun, "The homotopy method for identifying the radiative source term in the heat conduction problem," *Applicable Anal.*, vol. 95, no. 4, pp. 842–859, 2016. [Online]. Available: http://dx.doi.org/10.1080/00036811.2015.1037066

[19] P. del Aguila Pla and J. Jaldén, "Cell detection by functional inverse diffusion and non-negative group sparsity—Part I: Modeling and Inverse problems," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5407–5421, 2018.

[20] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *J. Roy. Statistical Soc., Series B (Statist. Methodology)*, vol. 68, no. 1, pp. 49–67, 2006. [Online]. Available: http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x

[21] G. Teschke, “Multi-frame representations in linear inverse problems with mixed multi-constraints,” *Appl. Comput. Harmonic Anal.*, vol. 22, no. 1, pp. 43–60, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1063520306000686

[22] M. Fornasier and H. Rauhut, “Recovery algorithms for vector-valued data with joint sparsity constraints,” *SIAM J. Numer. Anal.*, vol. 46, no. 2, pp. 577–613, 2008.

[23] 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. [Online]. Available: http://dx.doi.org/10.1137/080716542

[24] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York, NY, USA: Springer, 2011, ch. 27. Proximal minimization, pp. 399–413. [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-9467-7_27

[25] N. Parikh and S. Boyd, “Proximal algorithms,” *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014. [Online]. Available: http://dx.doi.org/10.1561/2400000003

[26] Y.-L. Yu, “On decomposing the proximal map,” in *Proc. Adv. Neural Inf. Process. Syst.,* 2013, pp. 91–99. [Online]. Available: http://papers.nips.cc/paper/4863-on-decomposing-the-proximal-map.pdf

[27] S. Bonettini and M. Prato, “New convergence results for the scaled gradient projection method,” *Inverse Problems*, vol. 31, no. 9, 2015, Art. no. 095008. [Online]. Available: http://stacks.iop.org/0266-5611/31/i=9/a=095008

[28] N. Pustelnik and L. Condat, “Proximity operator of a sum of functions; Application to depth map estimation,” *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1827–1831, Dec. 2017.

[29] P. L. Combettes and J. C. Pesquet, “A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 564–574, Dec. 2007.

[30] 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. [Online]. Available: https://doi.org/10.1137/080727749

[31] R. Gu and A. Dogandžić, “Projected Nesterov’s proximal-gradient algorithm for sparse signal recovery,” *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3510–3525, Jul. 2017.

[32] L. M. Briceño-Arias and P. L. Combettes, “Convex variational formulation with smooth coupling for multicomponent signal decomposition and recovery,” *Numer. Math. Theory Methods Appl.*, vol. 2, pp. 485–508, 2009.

[33] P. L. Combettes and C. L. Müller, “Perspective functions: Proximal calculus and applications in high-dimensional statistics,” *J. Math. Anal. Appl.*, vol. 457, no. 2, pp. 1283–1306, 2018, special Issue on Convex Analysis and Optimization: New Trends in Theory and Applications. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022247X16308071

[34] M. Yukawa and H. Kagami, “Supervised nonnegative matrix factorization via minimization of regularized Moreau-envelope of divergence function with application to music transcription,” *J. Franklin Inst.*, vol. 355, no. 4, pp. 2041–2066, 2018, Special issue on recent advances in machine learning for signal analysis and processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0016003217306336

[35] D. Chen and R. J. Plemmons, “Nonnegativity constraints in numerical analysis,” in *The Birth of Numerical Analysis*. Singapore: World Scientific, Nov. 2009, pp. 109–139.

[36] P. L. Combettes, A. M. McDonald, C. A. Micchelli, and M. Pontil, “Learning with optimal interpolation norms,” 2016, arXiv:1603.09273.

[37] D. G. Luenberger, *Optimization by Vector Space Methods*. Hoboken, NJ, USA: Wiley, 1969.

[38] 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. [Online]. Available: http://dx.doi.org/10.1007/s10957-015-0746-4

[39] A. Y. Karulin and P. V. Lehmann, *Handbook of ELISPOT: Methods and Protocols*, 2nd ed. New York, NY, USA: Springer, 2012, ch. 11, pp. 125–143.

[40] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000. [Online]. Available: https://doi.org/10.1023/A:1026543900054

[41] M. Grant and S. Boyd, “CVX: MATLAB software for disciplined convex programming, version 2.1,” Mar. 2014. [Online]. Available: http://cvxr.com/cvx

[42] M. ApS, *The MOSEK Optimization Toolbox for MATLAB Manual. Version 8.1.*, 2017. [Online]. Available: http://docs.mosek.com/8.1/toolbox/index.html

[43] L. Xiao and T. Zhang, “A proximal-gradient homotopy method for the sparse least-squares problem,” *SIAM J. Optim.*, vol. 23, no. 2, pp. 1062–1091, 2013.

[44] R. Eghbali and M. Fazel, “Decomposable norm minimization with proximal-gradient homotopy algorithm,” *Comput. Optim. Appl.*, vol. 66, no. 2, pp. 345–381, Mar. 2017. [Online]. Available: https://doi.org/10.1007/s10589-016-9871-8

[45] Z. Jia, X. Cai, and D. Han, “Comparison of several fast algorithms for projection onto an ellipsoid,” *J. Comput. Appl. Math.*, vol. 319, pp. 320–337, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377042717300122

Pol del Aguila Pla (S’15) received a double degree in telecommunications and electrical engineering from the Universitat Politècnica de Catalunya, Barcelona, Spain, and the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2014. Since August 2014, he is currently working toward the Ph.D. degree in electrical engineering and signal processing with KTH under the supervision of Joakim Jaldén. His Ph.D. work includes the research collaboration with Mabtech AB that led to the results published here and the development of the ELISPOT and Flourospot reader Mabtech IRIS^{TM}. Since August 2015, he is a Reviewer for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. During 2017, he received a number of grants to support the international promotion of his research in inverse problems for scientific imaging, including a 2017 KTH Opportunities Fund scholarship, a Knut and Alice Wallenberg Jubilee appropriation, an Åforsk Foundation’s scholarship for travel and a 2017 Engineering Sciences grant from the Swedish Academy of Sciences (KVA, ES2017-0011).

Joakim Jaldén (S’03–M’08–SM’13) received the M.Sc. and Ph.D. degrees in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2002 and 2007, respectively. From July 2007 to June 2009, he held a postdoctoral research position with the Vienna University of Technology, Vienna, Austria. He also studied at Stanford University, Stanford, CA, USA, from September 2000 to May 2002, and worked at ETH, Zürich, Switzerland, as a Visiting Researcher, from August to September, 2008. In July 2009, he returned to KTH, where he is currently a Professor of signal processing. His recent work includes work on signal processing for biomedical data analysis, and the automated tracking of (biological) cell migration and morphology in time-lapse microscopy in particular. Early work in this field was awarded a conference best paper Award at IEEE ISBI 2012, and subsequent work by the group has been awarded several Bitplane Awards in connection to the ISBI cell tracking challenges between 2013 and 2015. He was an Associate Editor for theIEEE COMMUNICATIONS LETTERS between 2009 and 2011, and an Associate Editor for the IEEE TRANSACTIONS IN SIGNAL PROCESSING between 2012 and 2016. Since 2013, he has been a member of the IEEE Signal Processing for Communications and Networking Technical Committee , where he is currently a Vice-Chair. Since 2016, he has also been responsible for the five year B.Sc and M.Sc. Degree Program in electrical engineering with KTH.

For his work on MIMO communications, he has been awarded the IEEE Signal Processing Societies 2006 Young Author Best Paper Award, the Distinguished Achievement Award of NEWCOM++ Network of Excellence in Telecommunications 2007–2011, and the best student conference paper Award at IEEE ICASSP 2007. He is also the recipient of the Ingvar Carlsson Career Award issued in 2009 by the Swedish Foundation for Strategic Research.