Minimization and Parameter Estimation for Seminorm Regularization Models with I-Divergence Constraints

T. Teuber[*], G. Steidl[*] and R. H. Chan[†]

November 26, 2012

Abstract

In this papers we analyze the minimization of seminorms $\|L \cdot \|$ on $\mathbb{R}^n$ under the constraint of a bounded I-divergence $D(b, H\cdot)$ for rather general linear operators H and L. The I-divergence is also known as Kullback-Leibler divergence and appears in many models in imaging science, in particular when dealing with Poisson data. Often H represents, e.g., a linear blur operator and L is some discrete derivative or frame analysis operator. We prove relations between the the parameters of I-divergence constrained and penalized problems without assuming the uniqueness of their minimizers. To solve the I-divergence constrained problem we apply first-order primal-dual algorithms which reduce the problem to the solution of certain proximal minimization problems in each iteration step. One of these proximation problems is an I-divergence constrained least squares problem which can be solved based on Morosov's discrepancy principle by a Newton method. Interestingly, the algorithm produces not only a sequence of vectors which converges to a minimizer of the constrained problem but also a sequence of parameters which convergences to a regularization parameter so that the corresponding penalized problem has the same solution as our constrained one. We demonstrate the performance of various algorithms for different image restoration tasks both for images corrupted by Poisson noise and multiplicative Gamma noise.

1 Introduction

Regularized ill-posed problems were rigorously investigated by mathematicians since the early 60s of the last century, see for example the seminal book [46] and the survey paper [41]. One of the best examined models in $\mathbb{R}^n$ is

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{\lambda}{2} \|b - Hx\|_2^2 + \|Lx\|_2^2 \right\}, \quad \lambda > 0, \tag{1}$$

where $b \in \mathbb{R}^n$ is the H-transformed and perturbed signal. The known linear transform operator $H \in \mathbb{R}^{n,n}$ is in general ill-conditioned or not invertible. The linear operator $L \in \mathbb{R}^{m,n}$ in the regularization term enforces some regularity of the minimizer. Examples are discrete derivative operators or nonlocal operators as considered in [39, 63]. A key issue of this model is the determination of a suitable regularization parameter λ, which balances the data fidelity with the regularity of the solution. Several techniques were developed to address this

[*]University of Kaiserslautern, Dept. of Mathematics, Kaiserslautern, Germany
[†]Chinese University of Hong Kong

topic, e.g., Morozov's discrepancy principle [51], the L-curve criterion [48], the generalized cross-validation [69], normalized cumulative or residual periodogram approaches [42, 58] and variational Bayes' approaches [2, 53]. In this paper, we will adapt the idea of the discrepancy principle which chooses the 'optimal' regularization parameter $\hat{\lambda}$ such that the norm of the defect $\|b - H\hat{x}(\hat{\lambda})\|_2^2$ between the corrupted data b and the minimizer $\hat{x}(\hat{\lambda})$ of (1) equals some known error τ, i.e., solves the *discrepancy equation* $\|b - H\hat{x}(\lambda)\|_2^2 = \tau$. To efficiently compute $\hat{\lambda}$ it is important that $\hat{x}(\lambda)$ is given by an explicit analytical expression.

When dealing with image processing applications the above model is often replaced by

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \lambda \Phi(b, Hx) + \|Lx\| \right\}, \qquad \lambda > 0 \tag{2}$$

with an appropriate data fidelity term $\Phi(b, H\cdot)$ and certain norms $\|\cdot\|$ on $\mathbb{R}^m$ to get an edge-preserving restoration model. Note that any seminorm on $\mathbb{R}^n$ can be written in the form $\|L\cdot\|$ with an appropriate linear operator $L \in \mathbb{R}^{m,n}$. The frequently applied approach of Rudin, Osher and Fatemi [57] involves for example $TV(x) := \| |\nabla x| \|_1$ as regularization term, where $L = \nabla$ denotes a discrete gradient operator and $\| |\cdot| \|_1$ the mixed ℓ_1-norm. Other models use discrete frame operators L together with the (mixed) ℓ_1-norm to enforce (group) sparsity of the frame coefficients.

Recently, also constrained optimization models were successfully applied in imaging and sparsity tasks, see, e.g., [18, 24, 27, 35, 52, 66, 70]. Such constrained models can have the advantage that some knowledge on the noise allows to estimate its parameter τ better than the regularization parameter λ of the penalized model. In particular,

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ TV(x) \quad \text{subject to} \quad \|Hx - b\|_2^2 \leq \tau \right\} \tag{3}$$

was considered in [52, 71], where the authors in [16] consider the problem from the point of view of the penalized problem (2). But rather than fixing λ in all iterations, they tune λ in each iteration step such that the corresponding parameter sequence converges to some optimal $\hat{\lambda}$ with the property that the minimizer $\hat{x}(\hat{\lambda})$ of the corresponding penalized problem fulfills the discrepancy equation.

In this paper, we are interested in the I-divergence $D(b, H\cdot)$ instead of the squared ℓ_2-norm $\|H \cdot -b\|_2^2$ as data fidelity term, which is more appropriate if the data is corrupted, e.g., by Poisson noise or multiplicative noise, see, e.g., [4, 14, 21, 47, 49, 63, 72]. Poisson data typically occurs in imaging processes where the images are obtained by counting particles, e.g., photons, that hit the image domain, see [7]. Multiplicative noise often appears as speckle in applications like laser, ultrasonic [15, 68] or synthetic aperture radar (SAR) imaging [13, 50]. We want to analyse the I-divergence constrained problem

$$\operatorname*{argmin}_{x \geq 0} \left\{ \|Lx\| \quad \text{subject to} \quad D(b, Hx) \leq \tau \right\} \tag{4}$$

under rather general assumptions on H and L and to examine the relation to the penalized problem

$$\operatorname*{argmin}_{x \geq 0} \left\{ \|Lx\| + \lambda D(b, Hx) \right\}. \tag{5}$$

Discrepancy principles for these problems were discussed in [3, 4, 8, 18], where it was proposed e.g., in [8] that, under certain conditions, the 'optimal' regularization parameter $\hat{\lambda}$ in (5)

should be the solution of the discrepancy equation

$$D(b, \hat{x}(\lambda)) = \tau \tag{6}$$

where $\tau := \frac{n}{2}$. However, since there is no explicit expression for the minimizer $\hat{x}(\lambda)$ of (5) the 'optimal' parameter $\hat{\lambda}$ cannot be efficiently computed from (6).

In this paper we use the same basic idea as recently proposed by Carlavan and Blanc-Feraud in [17, 18] for blur operators H and the frame operator L of the dual-tree complex wavelet transform [59]: First order primal-dual algorithms restrict the minimization of the constrained problem (16) to the iterative solution of certain simpler proximal minimization problems. One of these problems appears to be an I-divergence constrained least squares problem

$$\operatorname*{argmin}_{t \in \mathbb{R}^n} \left\{ \frac{1}{2} \|t - a\|_2^2 \quad \text{subject to} \quad D(b, t) \leq \tau \right\}.$$

The corresponding I-divergence penalized least squares problem

$$\operatorname*{argmin}_{t \in \mathbb{R}^n} \left\{ \frac{1}{2} \|t - a\|_2^2 + \lambda D(b, t) \right\}, \quad \lambda > 0$$

has an analytical solution $\hat{t}(a, \lambda)$ which can be used in the least squares - related discrepancy equation

$$D(b, \hat{t}(a, \lambda)) = \tau$$

to compute its solution efficiently by Newton's method. We show that such primal-dual algorithms compute beyond the minimizer $\hat{x}$ of (4) also the regularization parameter $\hat{\lambda}$ such that $\hat{x}$ fulfills the general discrepancy principle (6).

The structure of this paper is as follows: In Section 2 we provide the basic notation and recall useful relations. In particular, we provide a theorem on the general relation between constrained and penalized convex problems and specify this for I-divergence constrained and penalized least squares problems. Section 3 analyzes the constrained problem (4) and the penalized problem (5) under rather general assumptions on H and L. We will see that under such mild assumptions both problems have solutions and that different solutions of the same problem leave $\|L \cdot\|$ and $H \cdot$ fixed. In Section 4, we deal with the minimization of the seminorm - I-divergence constrained problem. First, we introduce the dual problems and consider their relations to the primal ones. Then, we apply primal-dual first order methods together with a discrepancy principle and a Newton algorithm for the appearing inner least squares - I-divergence problems. We prove that on the one hand this algorithm converges to a solution of (4) and on the other hand computes the regularization parameter $\hat{\lambda}$ which fulfills the discrepancy equation 6. In Section 5, we show how to choose appropriate parameters τ in the cases of multiplicative Gamma-distributed noise and recall a result on Poisson noise from [8]. Section 6 demonstrates the performance of our algorithms both for the denoising of images containing multiplicative Gamma-distributed noise and for deblurring images corrupted by Poisson noise. We finish the paper with conclusions in Section 7. The Appendix contains some auxiliary lemmas and provides standard relations on dual problems.

2 Notation and Basic Relations

2.1 Notation

In this paper we deal with functions $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. By $\mathrm{lev}_\tau \Phi := \{x : \Phi(x) \leq \tau\}$ we denote the (lower) *level sets* of Φ. For $x^* \in \mathbb{R}^n$, where $\Phi(x^*)$ is finite, the *subdifferential* $\partial \Phi(x^*)$ of Φ at x^* is the set

$$\partial \Phi(x^*) := \{p \in \mathbb{R}^n : \langle p, x - x^* \rangle \leq \Phi(x) - \Phi(x^*) \, \forall x \in \mathbb{R}^n\}.$$

If Φ is proper, convex and $x^* \in \mathrm{ri}(\mathrm{dom}\Phi)$, then $\partial \Phi(x^*) \neq \emptyset$. The Fenchel *conjugate function* of Φ is defined by

$$\Phi^*(p) := \sup_{x \in \mathbb{R}^n} \{\langle p, x \rangle - \Phi(x)\}.$$

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ with dual norm $\|\cdot\|_* := \max_{\|x\| \leq 1} \langle \cdot, x \rangle$. By $B_{\|\cdot\|}(r) := \{x \in \mathbb{R}^n : \|x\| \leq r\}$ we denote the ball with respect to $\|\cdot\|$ with center 0 and radius r. Further

$$\iota_S(x) := \begin{cases} 0 & \text{if } x \in S, \\ +\infty & \text{otherwise} \end{cases}$$

signifies the *indicator function* ι_S of a set $S \neq \emptyset$ and P_S the *orthogonal projector* onto S. For a norm we have

$$\partial \|x\| = \begin{cases} B_{\|\cdot\|_*}(1) & \text{if } \|x\| = 0, \\ \{p \in \mathbb{R}^n : \langle p, x \rangle = \|x\|, \|p\|_* = 1\} & \text{otherwise} \end{cases} \tag{7}$$

and

$$\|p\|^* = \iota_{\mathrm{lev}_1 \|\cdot\|_*}(p).$$

For the indicator function of a convex set $S \neq \emptyset$ it holds for $x \in S$ that $\partial \iota_S(x) = N_S(x)$, where N_S denotes the normal cone to S at $x \in S$ and $\iota_S^* = \sigma_S$ with the *support function* $\sigma_S(x) := \sup_{y \in S} \langle x, y \rangle$. Moreover, $\sigma_S^* = \iota_S$ if S is in addition closed. For $S := \mathbb{R}_{\geq 0}^n$ and $x \geq 0$, we have for example

$$\partial \iota_{\mathbb{R}_{\geq 0}^n}(x) = N_{\mathbb{R}_{\geq 0}^n}(x) = \mathcal{I}_1 \times \ldots \times \mathcal{I}_n, \quad \text{where} \quad \mathcal{I}_k := \begin{cases} (-\infty, 0] & \text{if } x_k = 0, \\ \{0\} & \text{if } x_k > 0 \end{cases} \tag{8}$$

and $\iota_{\mathbb{R}_{\geq 0}^n}^* = \sigma_{\mathbb{R}_{\geq 0}^n} = \iota_{\mathbb{R}_{\leq 0}^n}$.

For given $b \in \mathbb{R}_{>0}^n$ and 1_n denoting a vector consisting of n ones, the discrete *I-divergence* also known as *generalized Kullback-Leibler divergence* is defined by

$$D(b, t) := \begin{cases} \langle 1_n, b \log \frac{b}{t} - b + t \rangle & \text{if } t > 0, \\ +\infty & \text{otherwise,} \end{cases}$$

cf. [26]. Note that

$$D(b, t) = \langle 1_n, t - b \log t \rangle - \langle 1_n, b - b \log b \rangle \quad \text{for } t > 0.$$

Using the agreement that $0 \log 0 := 0$ it is possible to generalize the definition of the *I*-divergence to $b \geq 0$. In this paper, we restrict our attention to $b > 0$ since we will use

that in this case the function $D(b, \cdot)$ is strictly convex and has b as unique minimizer, where $D(b, b) = 0$. Since $D(b, \cdot)$ is proper, convex and continuous, the level sets

$$\mathrm{lev}_\tau D(b, \cdot) := \{t \in \mathbb{R}^n : D(b, t) \leq \tau\}$$

are convex and closed. Moreover, $\mathrm{lev}_\tau D(b, \cdot) \neq \emptyset$ if and only if $\tau \geq 0$. Using the agreement that $0 \log 0 := 0$ it is possible to generalize the definition of the I-divergence to $b \geq 0$. In this paper, we restrict our attention to $b > 0$. The conjugate function of $D(b, \cdot)$ is given by

$$D^*(b, p) := \begin{cases} -\langle b, \log(1_n - p) \rangle & \text{if } p < 1_n, \\ +\infty & \text{otherwise.} \end{cases}$$

2.2 Relation between penalized and constrained convex problems

We will use the following well-known relation between constrained and penalized convex problems, see, e.g., [45].

Theorem 2.1. *For proper, convex, lower semi-continuous functions $F, G : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, where F is continuous, the problems*

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \{G(x) + \lambda F(x)\}, \qquad \lambda \geq 0 \tag{9}$$

and

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \{G(x) \quad \text{subject to} \quad F(x) \leq \tau\} \tag{10}$$

are related as follows:
i) Assume that $\mathrm{dom}\, F \cap \mathrm{dom}\, G \neq \emptyset$. Let $\hat{x}$ be a minimizer of (9). If $\lambda = 0$, then $\hat{x}$ is also a minimizer of (10) if and only if $\tau \geq F(\hat{x})$. If $\lambda > 0$, then $\hat{x}$ is also a minimizer of (10) for $\tau := F(\hat{x})$. Moreover, this τ is unique if and only if $\hat{x}$ is not a minimizer of G.
ii) Assume that $\mathrm{ri}(\mathrm{lev}_\tau F) \cap \mathrm{ri}(\mathrm{dom}\, G) \neq \emptyset$. Let $\hat{x}$ be a minimizer of (10). If $\hat{x}$ is not a minimizer of F, then there exists a parameter $\lambda \geq 0$ such that $\hat{x}$ is also a minimizer of (9). If $\hat{x}$ is in addition not a minimizer of G, then $\lambda > 0$.

Concerning i) we mention that in case the minimizer of (9) is not unique, say $\hat{x}_1 \neq \hat{x}_2$, the relation $F(\hat{x}_1) \neq F(\hat{x}_2)$ can appear.
Concerning ii) note that there may exist in general various parameters λ corresponding to the same parameter τ. For examples we refer to [25]. For finite, convex penalizers (which does not include $D(\cdot, b)$), the existence of a Lagrange multiplier $\lambda \geq 0$ is assured by [55, Corollary 28.2.1] and it is known that the set of Lagrange multipliers λ is a bounded closed interval, see [55, Corollary 29.1.5].
In the next section, will apply Theorem 2.1 with respect to the functions $F := D(b, H\cdot)$ and $G := \|L \cdot \| + \iota_{\mathbb{R}^n_{\geq 0}}$. We will see that for appropriately chosen τ and a (not necessarily unique) solution $\hat{x}$ of (10) there exists a unique λ such that $\hat{x}$ is also a solution of (9). Moreover, we will see that this λ appears a a by-product of our algorithm.

2.3 Least Squares - I-Divergence Problems

The main part of our algorithms for solving (4) will consist in the solution of least squares problems with constrained I-divergence

$$\operatorname*{argmin}_{t \in \mathbb{R}^n}\{\frac{1}{2}\|t - a\|_2^2 \quad \text{subject to} \quad D(b, t) \leq \tau\}, \quad \tau \geq 0 \tag{11}$$

with $a \in \mathbb{R}^n$ by utilizing the corresponding penalized problems

$$\operatorname*{argmin}_{t \in \mathbb{R}^n}\{\frac{1}{2}\|t - a\|_2^2 + \lambda D(b, t)\}, \quad \lambda \geq 0. \tag{12}$$

Since the functionals are coercive and strictly convex, both problems have a unique solution. If $a = b > 0$, then the solution is given by $\hat{t} = a$ for all $\tau, \lambda \geq 0$. If $a \neq b$, we obtain the solution by the following lemma.

Lemma 2.2. *Let $a, b \in \mathbb{R}^n$ with $b > 0$ and $a \neq b$ be given.*
i) Let $\lambda = 0$. Then problem (12) has the solution $\hat{t} = a$. This is also a solution of (11) if and only if $a > 0$ and $\tau \geq D(b, a)$. For $\lambda > 0$ problem (12) has the solution

$$\hat{t}(a, \lambda) = \frac{1}{2}\left(a - \lambda + \sqrt{(a - \lambda)^2 + 4\lambda b}\right), \tag{13}$$

where the notation has to be understood componentwise. In particular, $\hat{t} \notin \{a, b\}$. The discrepancy function

$$f(\lambda) := D(b, \hat{t}(a, \lambda)) = \langle 1_n, \hat{t}(a, \lambda)\rangle - \langle b, \log \hat{t}(a, \lambda)\rangle - \langle 1_n, b - b \log b\rangle$$

is strictly decreasing and convex and $\hat{t}$ is also the solution of (11) exactly for $\tau = D(b, g(a, \lambda))$.
ii) Let $\tau = 0$. Then problem (11) has the solution $\hat{t} = b$ and there does not exist $\lambda \geq 0$ such that $\hat{t} = b$ is the solution of (12). Let $\tau > 0$. Then the unique solution $\hat{t} > 0$ of problem (11) has the following properties: If $a > 0$ and $D(b, a) \leq \tau$, then $\hat{t} = a$ and this is also the solution of (12) exactly for $\lambda = 0$. Otherwise $\hat{t} \notin \{a, b\}$ and there exists a unique $\lambda > 0$ such that $\hat{t}$ is also the solution of (12).

Parts of this lemma can be found in several papers. The explicit form (13) of the minimizer $\hat{t}$ of (12) is contained, e.g., in [4, 10] and the discrepancy function f and its derivative f', e.g. in [18]. More generally, it was proved in [8] that for strictly convex, coercive and differentiable functions $\lambda D(b, \cdot) + \Psi$, $\lambda > 0$, the minimizer $\hat{t}(\lambda)$ has the property that $D(b, \hat{t}(\lambda))$ and $\Psi(\hat{t}(\lambda))$ are, respectively, a decreasing and an increasing function of λ. Of course our least squares - I-divergence model fits into this setting. For the whole proof of this lemma we refer to our preprint [65], where the following formulas for the discrepancy function f were derived (componentwise) with the abbreviation $\hat{t} = \hat{t}(a, \cdot)$:

$$f'(\lambda) = -\frac{(\hat{t}'(\lambda))^2}{\hat{t}(\lambda)}\sqrt{(a - \lambda)^2 + 4\lambda b} < 0, \tag{14}$$

$$f''(\lambda) = \left(\frac{\hat{t}'(\lambda)}{\hat{t}(\lambda)}\right)^2 \frac{a + \lambda + 2\sqrt{(a - \lambda)^2 + 4\lambda b}}{\sqrt{(a - \lambda)^2 + 4\lambda b}} > 0,$$

where

$$\hat{t}'(\lambda) = \frac{1}{2}\left(-1 + \frac{\lambda - a + 2b}{\sqrt{(a-\lambda)^2 + 4\lambda b}}\right) = \frac{-\hat{t}(\lambda) + b}{\sqrt{(a-\lambda)^2 + 4\lambda b}} > 0$$

and the inequalities hold true for $a \neq b$. By Lemma 2.2 there exists a unique solution of

$$f(\lambda) = D(b, \hat{t}(a, \lambda)) = \tau$$

which can be efficiently computed by Newton's method using (14).

3 Seminorm - I-Divergence Problems

In the following, let $H \in \mathbb{R}^{n,n}$ be such that $\{Hx : x \geq 0\} \cap \mathbb{R}^n_{>0} \neq \emptyset$, i.e., we have for the cone

$$\mathcal{K} := \{x \in \mathbb{R}^n_{\geq 0} : Hx > 0\} \neq \emptyset.$$

This is for example fulfilled for all x in the nonnegative orthant if H has only nonnegative entries and contains no zero row. It guarantees that

$$\tau_0 := \min_{x \geq 0} D(b, Hx) \tag{15}$$

is finite. Note that $\inf_{x \geq 0} D(b, Hx)$ is indeed attained, i.e., $\operatorname{argmin}_{x \geq 0} D(b, Hx) \neq \emptyset$ as shown in Lemma A.1 in the appendix. If $b \in \{Hx : x \geq 0\}$, we obtain $\tau_0 = 0$. Otherwise, we have $\tau_0 > 0$. Besides, $\operatorname{lev}_\tau D(b, H\cdot) \neq \emptyset$ for $\tau \geq \tau_0$.

For $L \in \mathbb{R}^{m,n}$ we are now interested in solving the constrained minimization problem

$$(P_{1,\tau}) \quad \operatorname*{argmin}_{x \geq 0}\{\|Lx\| \quad \text{subject to} \quad D(b, Hx) \leq \tau\}, \quad \tau \geq \tau_0, \tag{16}$$

which is closely related to the penalized problem

$$(P_{2,\lambda}) \quad \operatorname*{argmin}_{x \geq 0}\{\|Lx\| + \lambda D(b, Hx)\}, \quad \lambda \geq 0. \tag{17}$$

Setting

$$\tau_L := \min_{x \in \mathcal{N}(L),\, x \geq 0} D(b, Hx) \tag{18}$$

it holds that $\tau_L = +\infty$ if L is for example invertible. In the following, we will assume that $\tau_0 < \tau_L$, i.e.,

$$\operatorname*{argmin}_{x \geq 0} D(b, Hx) \cap \mathcal{N}(L) = \emptyset.$$

Example 3.1. *In image restoration the minimizers of functions involving the TV seminorm and the I-divergence often lead to good results. In this case, $L = \nabla$ is a discrete gradient operator as (29) with $\mathcal{N}(L) = \{\alpha 1_n : \alpha \in \mathbb{R}\}$. Moreover, H is often a blur operator which has usually nonnegative entries, contains no zero row and fulfills the condition $H^* 1_n = 1_n$. In this case, we automatically have $\mathcal{K} \neq \emptyset$.*

The bound τ_L can here be obtained as follows: With (18) and the structure of $\mathcal{N}(L)$ we have to find the minimizer of the function $\alpha \mapsto D(b, \alpha h)$, $\alpha > 0$, where $h := H1_n$. Due to the

condition $H^*1_n = 1_n$, it holds that $\langle 1_n, h \rangle = n$. Setting the derivative with respect to α of the function

$$D(b, \alpha h) = \langle 1_n, \alpha h - b \log(\alpha h) \rangle - \langle 1_n, b - b \log b \rangle$$
$$= \alpha n - \langle 1_n, b \log(\alpha h) \rangle - \langle 1_n, b - b \log b \rangle$$

to zero we obtain

$$0 = n - \frac{\langle 1_n, b \rangle}{\alpha} \quad \Leftrightarrow \quad \alpha = \frac{\langle 1_n, b \rangle}{n}.$$

This is minimizer of the function $D(b, \cdot h)$, since its second derivative is larger than zero for $\alpha > 0$. Thus, we have

$$\alpha 1_n = \operatorname*{argmin}_{x \in \mathcal{N}(L),\, x \geq 0} D(b, Hx) \quad with \quad \alpha = \frac{\langle 1_n, b \rangle}{n}$$

and

$$\tau_L = D(b, \alpha h) = \alpha n - \langle 1_n, b \log(\alpha h) \rangle - \langle 1_n, b - b \log b \rangle$$
$$= -\langle b, \log(\alpha h) \rangle + \langle b, \log b \rangle$$
$$= \left\langle b, \log \left(\frac{n}{\langle 1_n, b \rangle} \frac{b}{h} \right) \right\rangle.$$

Next, let us see under which conditions it holds that $\tau_0 = \tau_L$. Since $\mathcal{K} \neq \emptyset$ and $D(b, H\cdot)$ is continuous on its domain, we know by Fermat's rule that $\hat{x} \in \operatorname{argmin}_{x \geq 0} D(b, Hx)$ if and only if $\hat{x} \geq 0$ and

$$0 \in \nabla D(b, H\cdot)(\hat{x}) + \partial \iota_{\mathbb{R}^n_{\geq 0}}(\hat{x}) = H^* \left(1_n - \frac{b}{H\hat{x}} \right) + N_{\mathbb{R}^n_{\geq 0}}(\hat{x}) \quad \Leftrightarrow \quad H^* \frac{b}{H\hat{x}} - 1_n \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x}).$$

Since $N_{\mathbb{R}^n_{\geq 0}}(x) = \{0\}$ for all $x > 0$, we can conclude with $\hat{x} = \alpha 1_n > 0$ that

$$\tau_0 = \tau_L \quad \Leftrightarrow \quad H^* \frac{b}{h} = \alpha 1_n.$$

If H is invertible, this is only possible if $b = \alpha h$.

The following theorem clarifies the existence of a minimizer of the above problems and some of its properties.

Theorem 3.2. Let $H \in \mathbb{R}^{n,n}$ be such that $\mathcal{K} \neq \emptyset$ and $L \in \mathbb{R}^{m,n}$ fulfill $\mathcal{N}(H) \cap \mathcal{N}(L) = \{0\}$. Then the following relations are valid:

i) The problems $(P_{1,\tau})$, $\tau \geq \tau_0$ in (16) and $(P_{2,\lambda})$, $\lambda \geq 0$ in (17) have a solution.

ii) If $\hat{x}, \tilde{x}$ are solutions of $(P_{2,\lambda})$ for $\lambda > 0$, then

$$\|L\hat{x}\| = \|L\tilde{x}\| \quad and \quad H\hat{x} = H\tilde{x}. \tag{19}$$

iii) Let in addition $\operatorname{argmin}_{x \geq 0} D(b, Hx) \cap \mathcal{N}(L) = \emptyset$ and $\tau_0 < \tau < \tau_L$. If $\hat{x}, \tilde{x}$ are solutions of $(P_{1,\tau})$, then (19) holds true with $D(b, H\hat{x}) = \tau$.

8

Note that (19) implies
$$D(b, H\hat{x}) = D(b, H\tilde{x}).$$

Proof. i) The assertion is a consequence of Lemma A.2 applied to the setting

$$\mathbb{R}^n = \mathcal{R}(H^*) \oplus \mathcal{N}(H) = \mathcal{R}(L^*) \oplus \mathcal{N}(L)$$

with $\mathcal{N}(H) \cap \mathcal{N}(L) = \{0\}$ and $G := \|L \cdot \|$, $g := G|_{\mathcal{R}(L^*)}$, $J := \iota_{\mathbb{R}_{\geq 0}^n}$ and F defined problem dependent below. Note that $\operatorname{dom} G = \mathbb{R}^n$ and g has nonempty and bounded level sets $\operatorname{lev}_\beta g$ for $\beta \geq 0$.

In case of problem $(P_{1,\tau})$ we use $F := \iota_{\operatorname{lev}_\tau D(b, H\cdot)}$ and $f := F|_{\mathcal{R}(H^*)}$. Since $\tau \geq \tau_0$, we have that $\operatorname{dom} F \cap \operatorname{dom} G \cap \operatorname{dom} J \neq \emptyset$. Clearly, $\operatorname{lev}_\alpha f$ is nonempty and bounded for $\alpha \geq 0$.

In case of problem $(P_{2,\lambda})$ with $\lambda = 0$ any $\hat{x} \in \mathcal{N}(L)$ with $x \geq 0$ is a solution. For $\lambda > 0$ we use $F := \lambda D(b, H\cdot)$ and $f := F|_{\mathcal{R}(H^*)}$. Since $\mathcal{K} \neq \emptyset$, we have that $\operatorname{dom} F \cap \operatorname{dom} G \cap \operatorname{dom} J \neq \emptyset$. Clearly, $\operatorname{lev}_\alpha f$ is nonempty and bounded for $\alpha \geq \tau_0$.

ii) This assertion is a direct consequence of Lemma A.3 with $F := D(b, H\cdot)$, $G := \|L \cdot \| + \iota_{\mathbb{R}_{\geq 0}^n}$ and $\mathbb{R}^n = \mathcal{R}(H^*) \oplus \mathcal{N}(H)$.

iii) For problem $(P_{1,\tau})$ the first relation in (19) is straightforward. Next, we prove that $D(b, H\hat{x}) = \tau$ for any solution $\hat{x}$ of $(P_{1,\tau})$. We know by [9, Proposition 4.7.2] that since $\operatorname{lev}_\tau D(b, H\cdot) \cap \mathbb{R}_{\geq 0}^n \neq \emptyset$ and $\|L \cdot \|$ is continuous on its domain $\mathbb{R}^n$, there exists $v \in \partial \|L \cdot \|(\hat{x})$ such that

$$\langle x - \hat{x}, v \rangle \geq 0 \quad \forall x \in \operatorname{lev}_\tau D(b, H\cdot) \text{ with } x \geq 0. \tag{20}$$

We have that $v = L^* \partial \|L\hat{x}\|$. Since $\tau < \tau_L$, we know that $\hat{x} \notin \mathcal{N}(L)$. Thus, by (7), $v = L^* \hat{p}$ for some $\hat{p} \in \mathbb{R}^m$ with $\|\hat{p}\|_* = 1$ and $\langle \hat{p}, L\hat{x} \rangle = \langle v, \hat{x} \rangle = \|L\hat{x}\| > 0$. Hence, there exists at least one index $i_0 \in \{1, \ldots, n\}$ such that $v_{i_0} > 0$ and $\hat{x}_{i_0} > 0$.

If $D(b, H\hat{x}) < \tau$, then we conclude by the continuity of $D(b, H\cdot)$ that there exists a neighborhood of $\hat{x}$ such that $D(b, Hx) < \tau$ for all x in this neighborhood. Since $\hat{x} \geq 0$, we obtain that for small enough $\eta > 0$ the vector $x = (x_1, \ldots, x_n)^\mathrm{T}$ with $x_i := \hat{x}_i - \eta v_i$ if $\hat{x}_i > 0$ and $x_i := 0$ otherwise, lies in this neighborhood and fulfills $x \geq 0$. Using this x in (20) we obtain $-\eta \sum_{i \in \mathcal{I}} v_i^2 \geq 0$, where $\mathcal{I} \subset \{1, \ldots, n\}$ denotes the set of indices with $\hat{x}_i > 0$. Since i_0 belongs to $\mathcal{I}$, this is a contradiction and consequently $D(b, H\hat{x}) = \tau$.

To see the second relation in (19) assume that there exist two solutions $\hat{x} = \hat{x}_1 + \hat{x}_0 \geq 0$ and $\tilde{x} = \tilde{x}_1 + \tilde{x}_0 \geq 0$ of $(P_{1,\tau})$ with $\hat{x}_1, \tilde{x}_1 \in \mathcal{R}(H^*)$, $\hat{x}_1 \neq \tilde{x}_1$ and $\hat{x}_0, \tilde{x}_0 \in \mathcal{N}(H)$. Let $x = \mu\hat{x} + (1 - \mu)\tilde{x}$, $\mu \in (0, 1)$, so that $x \geq 0$. Since $D(b, H\cdot)$ is strictly convex on $\mathcal{R}(H^*)$, we have $D(b, Hx) < \tau$. On the other hand, we obtain

$$\|Lx\| \leq \mu\|L\hat{x}\| + (1 - \mu)\|L\tilde{x}\| = \|L\hat{x}\|$$

so that x is also a minimizer of $(P_{1,\tau})$, which is impossible, since we know from the previous part of the proof that any minimizer has to fulfill $D(b, Hx) = \tau$. This completes the proof. $\square$

Remark. As observed in [8], if H has only nonnegative entries and contains no zero row, then if $\mathcal{N}(\mathcal{H})$ is nontrivial, it consists only of vectors having at least one negative component. Thus $\|Hx\|$ is coercive on the nonnegative orthant and part ii) of Theorem 3.2 follows immediately. Our assumption on H is more general and requires just the existence of a vector $x \geq 0$ for which $Hx > 0$ is fulfilled.

Lemma 3.3. *Let $H \in \mathbb{R}^{n,n}$ be such that $\mathcal{K} \neq \emptyset$, $L \in \mathbb{R}^{m,n}$ fulfill $\mathcal{N}(H) \cap \mathcal{N}(L) = \{0\}$ and $\operatorname{argmin}_{x \geq 0} D(b, Hx) \cap \mathcal{N}(L) = \emptyset$. Let $\hat{x}$ be a solution of $(P_{2,\lambda})$ in (17) with $D(b, H\hat{x}) \neq \tau_L$. Then $\hat{x} \notin \mathcal{N}(L)$ and*

$$\lambda = \frac{\|L\hat{x}\|}{\langle 1_n, b - H\hat{x} \rangle}.$$

Proof. Since $\mathcal{K} \neq \emptyset$, we obtain by Fermat's rule that $\hat{x} \in \operatorname{argmin}_{x \geq 0}\{\|Lx\| + \lambda D(b, Hx)\}$ if and only if $\hat{x} \geq 0$ and

$$0 \in \partial \left(\|L \cdot \| + \lambda D(b, H\cdot) + \iota_{\mathbb{R}^n_{\geq 0}} \right)(\hat{x}),$$

$$0 \in L^*\partial\|L\hat{x}\| + \lambda H^*\nabla D(b, H\hat{x}) + \partial\iota_{\mathbb{R}^n_{\geq 0}}(\hat{x}),$$

$$0 \in L^*\partial\|L\hat{x}\| + \lambda H^* \left(1_n - \frac{b}{H\hat{x}} \right) + N_{\mathbb{R}^n_{\geq 0}}(\hat{x}),$$

$$\lambda H^* \left(\frac{b}{H\hat{x}} - 1_n \right) \in L^*\partial\|L\hat{x}\| + N_{\mathbb{R}^n_{\geq 0}}(\hat{x}).$$

By (7) this is fulfilled if and only if

$$\lambda H^* \left(\frac{b}{H\hat{x}} - 1_n \right) = L^*\hat{p}_2 + \hat{p}_3$$

for some $\hat{p}_3 \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x})$ and $\hat{p}_2 \in \mathbb{R}^m$ with $\|\hat{p}_2\|_* = 1$, $\langle \hat{p}_2, L\hat{x} \rangle = \|L\hat{x}\| > 0$ if $L\hat{x} \neq 0$ and $\|\hat{p}_2\|_* \leq 1$ otherwise. This implies

$$\lambda \left\langle \frac{b - H\hat{x}}{H\hat{x}}, H\hat{x} \right\rangle = \lambda \langle b - H\hat{x}, 1_n \rangle = \langle L^*\hat{p}_2 + \hat{p}_3, \hat{x} \rangle = \langle \hat{p}_2, L\hat{x} \rangle + \langle \hat{p}_3, \hat{x} \rangle.$$

Since $\hat{p}_3 \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x})$, it holds by (8) that $\langle \hat{p}_3, \hat{x} \rangle = 0$. If $\hat{x} \notin \mathcal{N}(L)$, we thus obtain $\lambda = \frac{\|L\hat{x}\|}{\langle 1_n, b - H\hat{x} \rangle}$. If $\hat{x} \in \mathcal{N}(L)$, then $\hat{x}$ can only be a solution of $(P_{2,\lambda})$ if $\hat{x} \in \operatorname{argmin}_{x \in \mathcal{N}(L), x \geq 0} D(b, Hx)$. But then we have the contradiction $D(b, H\hat{x}) = \tau_L$. $\qquad \square$

Using the previous considerations we can prove the following theorem on the relation between solutions of $(P_{1,\tau})$ and $(P_{2,\lambda})$.

Theorem 3.4. *Let $H \in \mathbb{R}^{n,n}$ be such that $\mathcal{K} \neq \emptyset$, $L \in \mathbb{R}^{m,n}$ fulfill $\mathcal{N}(H) \cap \mathcal{N}(L) = \{0\}$ and $\mathcal{N}(L) \cap \operatorname{argmin}_{x \geq 0} D(b, H\cdot) = \emptyset$. If $\hat{x}$ is a solution of $(P_{1,\tau})$ in (16) with $\tau_0 < \tau < \tau_L$, then there exists a unique $\lambda > 0$ such that $\hat{x}$ is also a solution of $(P_{2,\lambda})$ in (17). Moreover, λ does not depend on the chosen solution of $(P_{1,\tau})$.*

Proof. Let $\hat{x}$ be a solution of $(P_{1,\tau})$ for $\tau_0 < \tau < \tau_L$. We want to apply Theorem 2.1ii) with $F := D(b, H\cdot)$ and $G := \|L \cdot \| + \iota_{\mathbb{R}^n_{\geq 0}}$. Since $\tau > \tau_0$, we have that $\operatorname{ri}(\operatorname{lev}_\tau F) \cap \operatorname{dom} G \neq \emptyset$, which replaces the regularity assumption in the theorem, since $\iota_{\mathbb{R}^n \geq 0}$ is a polyhedral function. Since $\tau < \tau_L$, we have that $\hat{x} \geq 0$ is not a minimizer of G, i.e., $\hat{x} \notin \mathcal{N}(L)$. Further, $\hat{x}$ is not a minimizer of F by the following argument: Assume that $\hat{x} \geq 0$ is a minimizer of $D(b, H\cdot)$. Since $D(b, H\cdot)$ is continuous and $\tau > \tau_0$, we obtain that $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ with $x_i = \hat{x}_i + \eta(0 - \hat{x}_i) = (1 - \eta)\hat{x}_i$ if $\hat{x}_i > 0$ and $x_i = 0$ otherwise also fulfills $D(b, Hx) \leq \tau$ for sufficiently small $\eta > 0$. But then we get the contradiction

$$G(x) = \|Lx\| + \iota_{R^n_{\geq 0}}(x) = (1 - \eta)\|L\hat{x}\| < \|L\hat{x}\| + \iota_{R^n_{\geq 0}}(\hat{x}) = G(\hat{x}).$$

Thus, by Theorem 2.1ii) there exists $\lambda > 0$ such that $\hat{x}$ is also a solution of $(P_{2,\lambda})$. By Lemma 3.3 this λ is uniquely determined and by Theorem 3.2iii) it does not depend on the chosen solution of $(P_{1,\tau})$. $\qquad\square$

For matrices H with nonnegative entries and no zero row the authors in [8] proposed to choose the regularization parameter $\hat{\lambda}$ as solution of the discrepancy equation

$$D(b, H\hat{x}(\lambda)) = \tau, \quad \tau = \frac{n}{2}, \tag{21}$$

where $\hat{x}(\lambda)$ is the solution of $(P_{2,\lambda})$ in (17) and the choice of τ is explained in Section 5. The uniqueness of the solution $\hat{x}(\lambda)$ of $(P_{2,\lambda})$ was mentioned as one of the 'key issues' for dealing with (21). The criterion (21) was modified in [18] for $b \geq 0$ containing zero components and a criterion similar to (21) based on a quadratic approximation to the Kullback-Leibler divergence was proposed in [3].

However, since the solution $\hat{x}(\lambda)$ of $(P_{2,\lambda})$ in (17) is not given in an analytical form, the computation of the parameter $\hat{\lambda}$ fulfilling (21) still requires the numerical solution of a sequence of penalized problems with different regularizers λ.

In the following section we will show that $\hat{\lambda}$ can be obtained as limit of a sequence $\{\hat{\lambda}^{(k)}\}_k$ generated by a primal-dual algorithm to solve the corresponding constrained problem. The main idea is that the discrepancy principle is now applied in each step of algorithm to a much simpler problem than $(P_{2,\lambda})$ namely to an I-divergence penalized least squares problem (12). By Lemma 2.2 this simple problem has an analytical solution such that its strictly monotone, convex discrepancy function can be given analytically and the discrepancy equation can by solved efficiently by Newton's method. Note that our approach does not require that the solution $\hat{x}(\lambda)$ of $(P_{2,\lambda})$ is unique.

4 Minimization of Seminorms with Constrained I-Divergence

In this section, we compute a solution of $(P_{1,\tau})$ for $\tau_0 < \tau < \tau_L$ by various primal-dual algorithms. First, we will apply an ADMM algorithm as in [17, 18] together with the discrepancy principle and the Newton algorithm to solve the appearing inner least squares problems with I-divergence constraints. We prove that on the one hand this algorithm converges to a solution of $(P_{1,\tau})$ and on the other hand computes the regularization parameter $\hat{\lambda}$ such that the penalized problem $(P_{2,\hat{\lambda}})$ has the same solution. Then, we discuss the application of other primal-dual algorithms.

To understand the structure of the algorithms we have to involve the dual problems of $(P_{1,\tau})$ and $(P_{2,\lambda})$, which will be done in the next subsection.

4.1 Primal and Dual Problems

To understand the structure of the algorithms we have to involve the dual problems of $(P_{1,\tau})$ and $(P_{2,\lambda})$. The problems $(P_{1,\tau})$ and $(P_{2,\lambda})$, $\lambda > 0$ can be rewritten as

$$(P_{1,\tau}) \quad \operatorname*{argmin}_{\substack{x \in \mathbb{R}^n \\ y \in \mathbb{R}^{2n+m}}} \Big\{ \underbrace{\langle 0, x \rangle}_{=: f_1(x)} + \underbrace{\iota_{\operatorname{lev}_\tau D(b,\cdot)}(y_1) + \|y_2\| + \iota_{y_3 \geq 0}(y_3)}_{:= f_2(y_1, y_2, y_3)} \text{ s.t. } \underbrace{\begin{pmatrix} H \\ L \\ I \end{pmatrix}}_{A} x = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \Big\}, \quad (22)$$

$$(P_{2,\lambda}) \quad \operatorname*{argmin}_{\substack{x \in \mathbb{R}^n \\ y \in \mathbb{R}^{2n+m}}} \Big\{ \langle 0, x \rangle + \lambda D(b, y_1) + \|y_2\| + \iota_{y_3 \geq 0}(y_3) \text{ s.t. } \begin{pmatrix} H \\ L \\ I \end{pmatrix} x = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \Big\}.$$

Using the duality relations in Appendix A.2, in particular (31), and the fact that $f_1^*(p) = 0$ for $p = 0$ and $f_1^*(p) = +\infty$ otherwise, we obtain that the dual problems of $(P_{1,\tau})$ and $(P_{2,\lambda})$, $\lambda > 0$, are given by

$$(D_{1,\tau}) \quad \operatorname*{argmin}_{p=(p_1^{\mathrm{T}}, p_2^{\mathrm{T}}, p_3^{\mathrm{T}})^{\mathrm{T}}} \{ \sigma_{\operatorname{lev}_\tau D(b,\cdot)}(p_1) + \iota_{\operatorname{lev}_1 \|\cdot\|_*}(p_2) + \iota_{\mathbb{R}^n_{\leq 0}}(p_3) \text{ s.t. } H^* p_1 + L^* p_2 + p_3 = 0 \},$$

$$(D_{2,\lambda}) \quad \operatorname*{argmin}_{p=(p_1^{\mathrm{T}}, p_2^{\mathrm{T}}, p_3^{\mathrm{T}})^{\mathrm{T}}} \Big\{ \lambda D^* \Big(b, \frac{p_1}{\lambda} \Big) + \iota_{\operatorname{lev}_1 \|\cdot\|_*}(p_2) + \iota_{\mathbb{R}^n_{\leq 0}}(p_3) \text{ s.t. } H^* p_1 + L^* p_2 + p_3 = 0 \Big\}.$$

Note that $\iota_{\operatorname{lev}_\tau D(b,\cdot)}(Hx) = \iota_{\operatorname{lev}_\tau D(b, H\cdot)}(x)$ and $H^* N_{\operatorname{lev}_\tau D(b,\cdot)} = N_{\operatorname{lev}_\tau D(b, H\cdot)}$.

The following theorem provides the Karush-Kuhn-Tucker optimality conditions and relates the solutions of the dual and primal problems. In the following, let $\operatorname{SOL}(X)$ denote the solution set of problem (X).

Lemma 4.1. *Let $H \in \mathbb{R}^{n,n}$ be such that $\mathcal{K} \neq \emptyset$ and $L \in \mathbb{R}^{m,n}$ such that $\mathcal{N}(H) \cap \mathcal{N}(L) = \{0\}$. Let $\tau > \tau_0$ and $\lambda > 0$. Then the following relations hold true:*

$$\left. \begin{array}{l} \hat{x} \in \operatorname{SOL}(P_{1,\tau}) \\ \hat{p} \in \operatorname{SOL}(D_{1,\tau}) \end{array} \right\} \quad \Leftrightarrow \quad \left\{ \begin{array}{c} \hat{p}_1 \in N_{\operatorname{lev}_\tau D(b,\cdot)}(H\hat{x}), \ \hat{p}_2 \in \partial\|L\hat{x}\|, \ \hat{p}_3 \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x}) \\ \text{such that} \quad H^* \hat{p}_1 + L^* \hat{p}_2 + \hat{p}_3 = 0, \end{array} \right. \quad (23)$$

and

$$\left. \begin{array}{l} \hat{x} \in \operatorname{SOL}(P_{2,\lambda}) \\ \hat{p} \in \operatorname{SOL}(D_{2,\lambda}) \end{array} \right\} \quad \Leftrightarrow \quad \left\{ \begin{array}{c} \hat{p}_1 = \lambda(1_n - \frac{b}{H\hat{x}}), \ \hat{p}_2 \in \partial\|L\hat{x}\|, \ \hat{p}_3 \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x}) \\ \text{such that} \quad H^* \hat{p}_1 + L^* \hat{p}_2 + \hat{p}_3 = 0. \end{array} \right. \quad (24)$$

Since $\operatorname{SOL}(P_{1,\tau})$ and $\operatorname{SOL}(P_{2,\lambda})$ are nonempty, the proof follows by standard arguments from the duality theory of convex functions, cf. [11].

The following subsections describe algorithms to solve $(P_{1,\tau})$.

4.2 ADMM Involving Least Squares Problems with I-Divergence Constraints

We apply the ADMM algorithm for solving $(P_{1,\tau})$ as in the PIDSplit+ algorithm in [61], see also [10, 36]. Considering $(P_{1,\tau})$ in the form (22) and solving the inner I-divergence constrained least square problems based on the discrepancy principle by a Newton method, we obtain the following algorithm which was recently also suggested in [17, 18].

Algorithm I (ADMM for solving $(P_{1,\tau})$)

Initialization: $q_1^{(0)} = q_2^{(0)} = q_3^{(0)} = 0$, $y_1^{(0)} = Hb$, $y_2^{(0)} = Lb$, $y_3^{(0)} = b$ and $\gamma > 0$.

For $k = 0, 1, \dots$ repeat until a stopping criterion is reached:

$$
\begin{aligned}
x^{(k+1)} &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \|q_1^{(k)} + Hx - y_1^{(k)}\|_2^2 + \|q_2^{(k)} + Lx - y_2^{(k)}\|_2^2 + \|q_3^{(k)} + x - y_3^{(k)}\|_2^2 \right\} \\
&= (H^{\mathrm{T}}H + L^{\mathrm{T}}L + I)^{-1} \left(H^{\mathrm{T}}(y_1^{(k)} - q_1^{(k)}) + L^{\mathrm{T}}(y_2^{(k)} - q_2^{(k)}) + (y_3^{(k)} - q_3^{(k)}) \right). \\
y_1^{(k+1)} &= \operatorname*{argmin}_{y_1 \in \mathbb{R}^n} \left\{ \iota_{\operatorname{lev}_\tau D(b, \cdot)}(y_1) + \frac{\gamma}{2} \| \underbrace{q_1^{(k)} + Hx^{(k+1)}}_{=:a^{(k+1)}} - y_1 \|_2^2 \right\}
\end{aligned}
$$

$$
= \begin{cases}
a^{(k+1)} & \text{if } a^{(k+1)} > 0 \text{ and } D(b, a^{(k+1)}) \leq \tau \\
\hat{t}\left(a^{(k+1)}, \frac{\lambda_{k+1}}{\gamma}\right) & \text{otherwise,}
\end{cases}
$$

$$
\lambda_{k+1} = \begin{cases}
0 & \text{if } a^{(k+1)} > 0 \text{ and } D(b, a^{(k+1)}) \leq \tau \\
\text{as solution of } D\left(b, \hat{t}\left(a^{(k+1)}, \frac{\lambda}{\gamma}\right)\right) = \tau & \text{otherwise.}
\end{cases}
$$

$$
\begin{aligned}
y_2^{(k+1)} &= \operatorname*{argmin}_{y_2 \in \mathbb{R}^m} \left\{ \|y_2\| + \frac{\gamma}{2}\|q_2^{(k)} + Lx^{(k+1)} - y_2\|_2^2 \right\} \\
&= \left(I - P_{B_{\|\cdot\|_*}(1/\gamma)} \right)\left(q_2^{(k)} + Lx^{(k+1)} \right) \\
y_3^{(k+1)} &= \operatorname*{argmin}_{y_3 \in \mathbb{R}^n} \left\{ \iota_{y_3 \geq 0}(y_3) + \frac{\gamma}{2}\|q_3^{(k)} + x^{(k+1)} - y_3\|_2^2 \right\}, \\
&= P_{\mathbb{R}_{\geq 0}}\left(q_3^{(k)} + x^{(k+1)} \right) \\
q_1^{(k+1)} &= q_1^{(k)} + Hx^{(k+1)} - y_1^{(k+1)}, \\
q_2^{(k+1)} &= q_2^{(k)} + Lx^{(k+1)} - y_2^{(k+1)}, \\
q_3^{(k+1)} &= q_3^{(k)} + x^{(k+1)} - y_3^{(k+1)}.
\end{aligned}
$$

The computation of $y_1^{(k+1)}$ is justified by Lemma 2.2, where $\hat{t}$ is given by (13). The unique solution λ_{k+1} of the discrepancy equation $D\left(b, \hat{t}\left(a^{(k+1)}, \frac{\lambda}{\gamma}\right)\right) = \tau$ can be computed by Newton's method.

The orthogonal projection onto $B_{\|\cdot\|_*}(1/\gamma)$ required in the computation of $y_2^{(k+1)}$ can be easily computed for the ℓ_p-norms with $p = 1, \infty$ and their mixed versions, see, e.g., [30, 62, 73].

Finally note that this is a so-called *scaled ADMM algorithm* where $q = p/\gamma$ replaces the dual variable p.

The convergence of the algorithm is ensured by the following theorem. In particular, we obtain that the sequence $\{\lambda_k\}_k$ converges to the regularization parameter $\hat{\lambda} > 0$ such that $\hat{x} = \lim_{k \to \infty} x^{(k)}$ is both a solution of $(P_{1,\tau})$ and of $(P_{2,\hat{\lambda}})$.

Theorem 4.2. *Let $b \in \mathbb{R}^n$, $b > 0$ and $L \in \mathbb{R}^{m,n}$, $H \in \mathbb{R}^{n,n}$ such that $\mathcal{N}(L) \cap \mathcal{N}(H) = \{0\}$ and $\operatorname{argmin}_{x \geq 0} D(b, Hx) \cap \mathcal{N}(L) = \emptyset$. Let $\tau_0 < \tau < \tau_L$. Then the sequence $\{(x^{(k)}, y^{(k)}, q^{(k)}, \lambda_k)\}_k$ generated by the ADMM Algorithm I converges to $(\hat{x}, \hat{y}, \hat{q}, \hat{\lambda})$, where $\hat{x}$ is a solution of problem $(P_{1,\tau})$ in (16) and of problem $(P_{2,\hat{\lambda}})$, $\hat{\lambda} > 0$ in (17) and $\hat{p} = \gamma\hat{q}$ is a solution of the dual problems $(D_{1,\tau})$ and $(D_{2,\hat{\lambda}})$. Further, $\hat{y} = (H^T L^T I)^T \hat{x}$ holds true.*

Proof. 1. The convergence of $\{(x^{(k)}, y^{(k)}, q^{(k)})\}_k$ to $(\hat{x}, \hat{y}, \hat{q})$, where $\hat{x} \in \mathrm{SOL}(P_{1,\tau})$, $\hat{p} = \gamma\hat{q} \in \mathrm{SOL}(D_{1,\tau})$ and $\hat{y} = (H^\mathrm{T}\, L^\mathrm{T}\, I)^\mathrm{T}\hat{x}$ follows from general convergence results of the ADMM, see, e.g., [33, 37, 60].

2. It remains to prove the convergence of $\{\lambda_k\}_k$. By part 1 of the proof we have that $a^{(k)} = Hx^{(k)} + q_1^{(k-1)}$ converges to $\hat{a} = \hat{y}_1 + \hat{q}_1$ and that $g(a^{(k)}, \frac{\lambda_k}{\gamma})$ converges to $\hat{y}_1$. Furthermore, it follows by componentwise computation that

$$g\left(a^{(k)}, \frac{\lambda_k}{\gamma}\right) = \frac{1}{2}\left(a^{(k)} - \frac{1}{\gamma}\lambda_k + \sqrt{\left(a^{(k)} - \frac{1}{\gamma}\lambda_k\right)^2 + 4\frac{1}{\gamma}b\,\lambda_k}\right) = y_1^{(k)},$$

$$\Leftrightarrow \sqrt{\left(a^{(k)} - \frac{1}{\gamma}\lambda_k\right)^2 + 4\frac{1}{\gamma}b\,\lambda_k} = 2y_1^{(k)} - \left(a^{(k)} - \frac{1}{\gamma}\lambda_k\right),$$

$$\Leftrightarrow \frac{1}{\gamma}\lambda_k(b - y_1^{(k)}) = y_1^{(k)}(y_1^{(k)} - a^{(k)}),$$

$$\Leftrightarrow \lambda_k(b - y_1^{(k)}) = -y_1^{(k)}p_1^{(k)}, \quad p_1^{(k)} := \gamma q_1^{(k)}. \tag{25}$$

Note that $g(a^{(k)}, 0) = a^{(k)}$, $a^{(k)} > 0$ is also contained in this setting. By Theorem 3.2iii) we know that $b - H\hat{x} = b - \hat{y}_1 \neq 0$, i.e., $b_i - \hat{y}_{1,i} \neq 0$ at least for one index $i \in \{1, \dots, n\}$. Thus, we see in the ith equation in (25) that $\lambda_k \to \hat{\lambda} = -\hat{y}_{1,i}\,\hat{p}_{1,i}/(b_i - \hat{y}_{1,i})$ as $k \to \infty$. Now, (23) implies that $\hat{p}_2 = \gamma\hat{q}_2 \in \partial\|L\hat{x}\|$ and $\hat{p}_3 = \gamma\hat{q}_3 \in N_{\mathbb{R}_{\geq 0}^n}(\hat{x})$ with $H^*\hat{p}_1 + L^*\hat{p}_2 + \hat{p}_3 = 0$. Moreover, we have by (25) and since $H\hat{x} > 0$ that

$$\hat{\lambda}\,(b - H\hat{x}) = -(H\hat{x})\,\hat{p}_1,$$

$$\hat{\lambda}\,(1_n - \frac{b}{H\hat{x}}) = \hat{p}_1.$$

Since $\tau < \tau_L$, it holds that $\hat{x} \notin \mathcal{N}(L)$. Hence, $\hat{\lambda} = 0$ would imply $\hat{p}_1 = 0$ and thus further $0 = L^*\hat{p}_2 + \hat{p}_3$ with $\|\hat{p}_2\|_* = 1$, $\langle\hat{p}_2, L\hat{x}\rangle = \|L\hat{x}\| > 0$. But then $0 = \langle\hat{x}, L^*\hat{p}_2 + \hat{p}_3\rangle$ and with (8) we have $0 = \langle L\hat{x}, \hat{p}_2\rangle = \|L\hat{x}\|$, which yields a contradiction. Consequently, $\hat{\lambda} > 0$ and $\hat{x}$, $\hat{p}$ fulfill the right-hand of (24). Therefore, they are also solutions of $(P_{2,\hat{\lambda}})$ and $(D_{2,\hat{\lambda}})$, respectively. $\qquad\square$

4.3 Other Primal-Dual Algorithms

Finally, we want to comment on other algorithms to solve $(P_{1,\tau})$. In particular, these algorithms avoid solving the linear system of equations in the computation of $x^{(k+1)}$. We emphasize that the purpose of this paper is not to compare different algorithms, but to show that our idea can be incorporated into several existing techniques.

Let us start with the Arrow-Hurwitz method [1], which was first used in image processing (with some speedup suggestions) in [75] under the name primal-dual hybrid gradient algorithm (PDHG). In general this algorithm computes a solution of

$$\operatorname*{argmin}_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \{f_1(x) + f_2(y) \quad \text{subject to} \quad Ax = y\}$$

as follows:

Algorithm (Arrow-Hurwitz Method, PDHG)
Initialization: $x^{(0)} = 0$, $p^{(0)} = 0$ and $s, t > 0$ with $st < \frac{1}{\|A\|_2^2}$.

For $k = 0, 1, \ldots$ repeat until a stopping criterion is reached:

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f_1(x) + \langle p^{(k)}, Ax \rangle + \frac{1}{2s} \|x - x^{(k)}\|_2^2 \right\}$$

$$= \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|x - (x^{(k)} - sA^*p^{(k)})\|_2^2 + sf_1(x) \right\},$$

$$p^{(k+1)} = \operatorname*{argmin}_{p \in \mathbb{R}^d} \left\{ f_2^*(p) - \langle p, Ax^{(k+1)} \rangle + \frac{1}{2t} \|p - p^{(k)}\|_2^2 \right\}$$

$$= \operatorname*{argmin}_{p \in \mathbb{R}^d} \left\{ \frac{1}{2} \|p - (p^{(k)} + tAx^{(k+1)})\|_2^2 + tf_2^*(p) \right\}.$$

For our setting (22) with $f_1 = 0$ the first step results in $x^{(k+1)} = x^{(k)} - sA^*p^{(k)}$. The second step of the algorithm can be decoupled into two parts, see [19, 75]:

$$y^{(k+1)} = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ f_2(y) + \frac{t}{2} \|\frac{1}{t} p^{(k)} + Ax^{(k+1)} - y\|_2^2 \right\}, \tag{26}$$

$$p^{(k+1)} = p^{(k)} + t(Ax^{(k+1)} - y^{(k+1)}). \tag{27}$$

For f_2 as in our setting (22) and $q^{(k)} := p^{(k)}/t$ these two steps are exactly those of the ADMM algorithm for updating $y = (y_1^{\mathrm{T}}, y_2^{\mathrm{T}}, y_3^{\mathrm{T}})^{\mathrm{T}}$ and $q = (q_1^{\mathrm{T}}, q_2^{\mathrm{T}}, q_3^{\mathrm{T}})^{\mathrm{T}}$, where we have to replace γ by t now. The Arrow-Hurwitz method was improved by involving an extrapolation step by Pock et al. in [54]. The convergence of the algorithm was proved in [19] (with some speedup suggestions). Using this extrapolation idea for the dual variable in its simplest form, the first step of the algorithm becomes

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|x - (x^{(k)} - sA^*(2p^{(k)} - p^{(k-1)}))\|_2^2 + sf_1(x) \right\}.$$

We summarize the algorithm which we call PDHGMp (PDHG with modified dual variable p) for our special setting:

Algorithm II (PDHGMp with inner Newton iterations)
Initialization: $(y^{(0)}) = \left((y_1^{(0)})^{\mathrm{T}}, (y_2^{(0)})^{\mathrm{T}}, (y_3^{(0)})^{\mathrm{T}} \right)^{\mathrm{T}}$ with $y_1^{(0)} = Hb$, $y_2^{(0)} = Lb$, $y_3^{(0)} = b$ and $q^{(0)} = 0$ and $s, t > 0$ with $st < \frac{1}{\|(H^{\mathrm{T}} L^{\mathrm{T}} I)\|_2^2}$.
For $k = 0, 1, \ldots$ repeat until a stopping criterion is reached:

$$x^{(k+1)} = x^{(k)} - st(H^{\mathrm{T}} L^{\mathrm{T}} I)(2q^{(k)} - q^{(k-1)}),$$
$$y^{(k+1)} \quad \text{as in Algorithm I with } \gamma := t,$$
$$q^{(k+1)} \quad \text{as in Algorithm I}.$$

Another algorithm for solving problem (2)/(3) was proposed in [71]. It is based on [16] and resembles in some way the dual method in [38]. The method in [38] uses a predictor-corrector scheme [23] in the alternating direction iterations for the dual variable. This algorithm can be adapted to our setting (22) as follows:

Algorithm (PDHG with Predictor-Corrector Step)
Initialization: $x^{(0)} = 0$, $p^{(0)} = 0$ and $s, t > 0$ with $st < \frac{1}{2\|A\|_2^2}$
For $k = 0, 1, \ldots$ repeat until a stopping criterion is reached:

$$p^{(k+\frac{1}{2})} = \operatorname*{argmin}_{p \in \mathbb{R}^d} \left\{ \frac{1}{2} \|p - (p^{(k)} + tAx^{(k)})\|_2^2 + tf_2^*(p) \right\}$$

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|x - (x^{(k)} - sA^*p^{(k+\frac{1}{2})})\|_2^2 + sf_1(x) \right\},$$

$$p^{(k+1)} = \operatorname*{argmin}_{p \in \mathbb{R}^d} \left\{ \frac{1}{2} \|p - (p^{(k)} + tAx^{(k+1)})\|_2^2 + tf_2^*(p) \right\}.$$

Note that the update steps for p can be splitted again as in (26)-(27).

This algorithm is efficient in the special case when $H = I$ is the identity matrix, e.g., for denoising problems in imaging. Instead of (22) the simpler constraint problem

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \{ \|Lx\| \quad \text{subject to} \quad D(b, x) \leq \tau \}, \quad \tau > 0 \tag{28}$$

has to be solved, which can be rewritten as

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \Big\{ \underbrace{\iota_{\mathrm{lev}_\tau D(b, \cdot)}(x)}_{f_1(x)} + \underbrace{\|y\|}_{f_2(y)} \quad \text{subject to} \quad Lx = y \Big\}.$$

Using that $f_2^*(p) = \iota_{\mathrm{lev}_1 \|\cdot\|_*}(p)$ the above algorithm becomes

Algorithm III (ADM with predictor-corrector step for minimizing (28))
Initialization: $x^{(0)} = b$, $p^{(0)} = Lb$, $\lambda_0 = 0$, $s, t > 0$ with $st < \frac{1}{2\|L\|_2^2}$.
For $k = 0, 1, \ldots$ repeat until a stopping criterion is reached:

$$p^{(k+\frac{1}{2})} = P_{B_{\|\cdot\|_*}(1)}(p^{(k)} + t\, Lx^{(k)}),$$

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|x - (x^{(k)} - sL^{\mathrm{T}}p^{(k+\frac{1}{2})})\|_2^2 \quad \text{subject to} \quad D(b, x) \leq \tau \right\},$$

$$p^{(k+1)} = P_{B_{\|\cdot\|_*}(1)}(p^{(k)} + t\, Lx^{(k+1)}).$$

The update step for the primal variable x requires again the solution of a least squares problem with I-divergence constraints, which can be done by Lemma 2.2 as follows:

$h^{(k+1)} = x^{(k)} - s\, L^{\mathrm{T}}p^{(k+\frac{1}{2})}$,

If $h^{(k+1)} > 0$ and $D(b, h^{(k+1)}) \leq \tau$, then

    $\lambda_{k+1} = 0$,

    $x^{(k+1)} = h^{(k+1)}$,

Otherwise

    Find λ_{k+1} as solution of $D(b, g(h^{(k+1)}, s\lambda)) = \tau$ by Newton's method initialized by λ_k,

    $x^{(k+1)} = g(h^{(k+1)}, s\lambda_{k+1})$.

A convergence proof of the algorithms can be given similarly to [71]. Note that one can prove convergence under the milder assumption $st < 1/\|A\|_2$ by using the technique in [19].

5 Choosing a Suitable Value for τ

As already pointed out in the introduction problems of the form (16) or rather (17) have been studied in the literature for the removal of Poisson or multiplicative Gamma noise in images, respectively, cf., [4, 47, 49, 63]. Here, it is typically assumed that $x \geq 0$ represents the original image vector and b is a corrupted version of x, which possibly underwent some linear transformation modeled by H and Hx is either corrupted by Poisson or multiplicative Gamma noise. To obtain a good reconstruction $\hat{x}$ of the original, noise-free image vector by (16) or (17), respectively, suitable values for λ and τ need to be chosen. In contrast to the regularization parameter λ in (17) a reasonable value for τ in (16) can usually be directly determined by statistical considerations if the type of noise corrupting the data is known. For Poisson data this was done e.g. in [3, 4, 8] and we recall the results from [8] in the following Lemma 5.1i) and Theorem 5.2i) while the parts ii) deal with multiplicative Gamma noise.
To explain the idea, let us first consider only one noisy pixel $b_i > 0$. Since this pixel is supposed to be corrupted by noise, it can be viewed as one realization of a random variable B_t with the given noise statistics. To determine now a reasonable value τ we may assume for a moment that the noise-free value $t = (Hx)_i$ of b_i is known and we may ask what mean value we can expect for our I-divergence term $D(b_i, t)$ for different noisy realizations b_i of B_t:

Lemma 5.1.   i) *Let B_t be a Poisson distributed random variable with expectation value $\mathbb{E}(B_t) = t > 0$. For t large enough it holds that*

$$\mathbb{E}\Big(B_t \log \frac{B_t}{t} - B_t + t\Big) = \frac{1}{2} + O\Big(\frac{1}{t}\Big).$$

ii) *Let V be a Gamma distributed random variable with density*

$$p_V(v) = \frac{K^K}{\Gamma(K)} v^{K-1} \exp(-K\,v)\, 1_{v\geq 0}(v), \qquad K \geq 1$$

and set $B_t := tV$. Then, we have

$$\mathbb{E}\Big(B_t \log \frac{B_t}{t} - B_t + t\Big) = t\big(\psi(K+1) - \log(K)\big),$$

where $\psi(x) := \frac{\partial}{\partial x} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ represents the digamma function and $\Gamma(x) := \int_0^\infty \exp(-s)\, s^{x-1}\, ds$ denotes the gamma function.

Proof. The proof of i) can be found in [74]. To prove ii) we use the definition of B_t and the fact that $\mathbb{E}(V) = 1$ so that

$$\begin{aligned}
\mathbb{E}\Big(B_t \log \frac{B_t}{t} - B_t + t\Big) &= \mathbb{E}\Big(t V \log V - t V + t\Big) \\
&= t\big(\mathbb{E}(V \log V) - \mathbb{E}(V) + 1\big) \\
&= t\,\mathbb{E}(V \log V).
\end{aligned}$$

Further, we obtain that

$$\mathbb{E}(V \log V) = \frac{K^K}{\Gamma(K)} \int\limits_0^\infty v^k \log v \, \exp(-Kv) \, dv$$

$$= \psi(K+1) - \log(K)$$

$(\Psi(x) = \int_0^\infty \frac{\exp(-s)}{s} - \frac{\exp(-xs)}{1-\exp(-s)} \, ds)$ so that finally,

$$\mathbb{E}\Big(B_t \log \frac{B_t}{t} - B_t + t\Big) = t\big(\psi(K+1) - \log(K)\big).$$

$\square$

Summing these results up over the whole image vectors we immediately obtain the following theorem:

Theorem 5.2. *Let $B = (B_1, \ldots, B_n)$ be a random vector and $t = (t_1, \ldots, t_n) \in \mathbb{R}_{>0}^n$.*

i) *If each B_i is Poisson distributed with expectation value t_i for $i = 1, \ldots, n$, then it holds that*

$$\mathbb{E}(\, D(B,t)\,) = \frac{1}{2} n + \sum_{i=1}^n O\Big(\frac{1}{t_i}\Big).$$

ii) *If all V_i are Gamma distributed and $B_i := t_i V_i$ for $i = 1, \ldots, n$, we have*

$$\mathbb{E}(\, D(B,t)\,) = \Big(\sum_{i=1}^n t_i\Big)\big(\psi(K+1) - \log(K)\big) = \Big(\sum_{i=1}^n \mathbb{E}(B_i)\Big)\big(\psi(K+1) - \log(K)\big).$$

This result shows that in case of Poisson noise and pixels with high original intensities t_i the expectation value of $D(B,t)$ is approximately $\frac{1}{2} n$ and thus, $\tau = \frac{1}{2} n$ is a good choice in (16). On the other hand, if the given image is corrupted by multiplicative Gamma noise, case ii) shows that

$$\tau = \Big(\sum_{i=1}^n \mathbb{E}(B_i)\Big)\big(\psi(K+1) - \log(K)\big)$$

is a reasonable choice, where $\sum_{i=1}^n \mathbb{E}(B_i)$ can well be approximated by $\sum_{i=1}^n b_i$. The following remark outlines the range of values τ we can expect for varying K:

Remark 5.3. *Using standard results for the digamma function ψ, see, e.g., [40, Sec. 8.36], it is not hard to show for case ii) that*

- $\mathbb{E}(\, D(B,t)\,)$ *is a strictly decreasing function in K $(K \geq 1)$,*

- *for $K = 1$ we have*

$$\mathbb{E}(\, D(B,t)\,) = (1-c)\Big(\sum_{i=1}^n \mathbb{E}(B_i)\Big) \approx 0.423\,\Big(\sum_{i=1}^n \mathbb{E}(B_i)\Big),$$

  where $c = 0,577...$ denotes the Euler-Mascheroni constant,

- $\mathbb{E}(\, D(B,t)\,) \to 0$ *for $K \to \infty$.*

6 Numerical Examples

Next, we want to illustrate the theoretical results of the former sections by numerical experiments with images corrupted by Poisson and multiplicative Gamma noise, respectively. For this purpose, we use for $\|Lx\|$ the mixed l_1-norm $\| \, | \cdot | \, \|_1$ and set L to be either a matrix modeling non-local similarities, see Remark 6.1, or the discrete gradient operator

$$\nabla := \left(\begin{array}{c} I \otimes D \\ D \otimes I \end{array} \right), \quad D := \begin{pmatrix} -1 & 1 & & & \\ 0 & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & 0 \end{pmatrix} \tag{29}$$

with $\otimes$ denoting the tensor product (Kronecker product) of matrices. In the latter case, $\|Lx\|$ becomes the discrete total variation $TV(x) := \| \, |\nabla x| \, \|_1$ mentioned in the introduction. We apply the peak signal to noise ratio (PSNR) defined by

$$\text{PSNR} \; = \; 10 \log_{10} \frac{|\max x_0 - \min x_0|^2}{\frac{1}{N} \|x - x_0\|_2^2}$$

for a quantitative comparison of the images x, where x_0 denotes the original image which we want to reconstruct. For solving problem (16) all algorithms are implemented in MATLAB and executed on a computer with an Intel Core i7-870 Processor (8M Cache, 2.93 GHz) and 8 GB physical memory.

6.1 Deblurring Facing Poisson Noise

Figure 1: *Left:* Original image with values scaled to [0, 3000] so that the brightest pixels correspond to 3000 detected photons. *Middle:* Corrupted image blurred by a Gaussian kernel (standard deviation 1.3) and contaminated by Poisson noise. *Right:* Restoration result by the I-divergence constrained model (16) with total variation seminorm.

Our first test image in Figure 1 shows a part of the 'cameraman' image, which has been corrupted by a Gaussian blur and contaminated by Poisson noise. The image gray values are here interpreted as photon counts in the range [0, 3000]. For synthetically adding Poisson
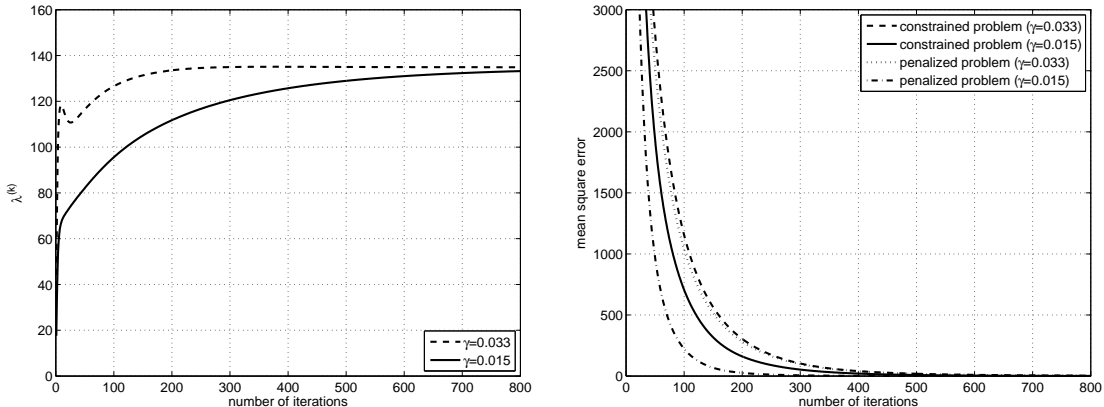
Figure 2: Convergence speed of $x^{(k)}$ and $\lambda^{(k)}$ in Algorithm I when computing the restored image in Figure 1 (right). *Left:* Iterates $\lambda^{(k)}$ for different parameters γ. *Right:* Evolution of the mean square errors $\frac{1}{N}\|x^{(k)} - x^*\|_2^2$ between the intermediate results $x^{(k)}$ and a sufficiently converged reference result x^*.

noise to the noise-free image we applied the MATLAB routine `imnoise(X,'poisson')`. This procedure assumes for data given in double precision that the input image X consists of the number of detected photons divided by 10^{12} - the maximal number of detectable photons. Therefore, we divided our given image by 10^{12} before applying this procedure and afterwards we scaled back again.

Computing the usually unknown value $D(b, Hx)$ for these test images yields a value of $0.5046\,n$, which is close to the estimate $\tau = 0.5\,n$ derived in Section 5. To restore the corrupted image we now solve the constrained minimization problem (16) with the total variation seminorm and $\tau = 0.5\,n$, which yields the good reconstruction depicted in Figure 1 (right). The minimization is here performed by the ADMM Algorithm I. As a by-product of the algorithm we obtain by Theorem 4.2 that the penalized problem (17) yields the same solution for a regularization parameter of $\lambda = 134.9$. As illustrated in Figure 2 the convergence speed of the iterates $x^{(k)}$ and $\lambda^{(k)}$ depends as usual on the chosen parameter $\gamma > 0$. Compared to a simplified version of Algorithm I for the penalized problem with fixed λ we see on the right that for our constrained problem Algorithm I is only slightly slower for equal values of γ.

The restoration result for the whole 'cameraman' image is shown in Figure 3. ADMM Algorithm I requires approximately 13.8 seconds (816 iterations) with optimized $\gamma = 0.0111$ to compute a solution $x^{(k)}$ so that the maximal relative pixel error $\|x^{(k)} - x^*\|_\infty/3000$ to a sufficiently converged reference result x^* is smaller than 1%. Here, the occurring linear system of equations has been solved using the discrete cosine II transform. For the same accuracy the PDHGMp Algorithm II requires only approximately 9 seconds (864 iterations) with $s = 16.9$ and $t = 0.01$. However, two parameters have to be optimized, here.

6.2 Denoising Facing Multiplicative Gamma Noise

TV Regularization Our next example in Figure 4 shows a 512×512 aerial image corrupted by multiplicative Gamma noise. The obtained restoration result by solving the constrained problem (16) with $H := I$, total variation seminorm and $\tau = \left(\sum_{i=1}^{n} b_i\right)\left(\psi(K+1) - \log(K)\right) \approx$

20

Figure 3: Result for the whole 'cameraman' image. *Left:* Original image of size 256×256 with values scaled to $[0, 3000]$. *Middle:* Corrupted image. *Right:* Restoration result by the *I*-divergence constrained model (16) with total variation seminorm.

$2.64\,n$ is depicted on the right. For computing this solution we used again Algorithm I. The occurring linear system of equations is solved here by a CG method, since due to the sparsity of the relevant matrices the CG method turned out to be even faster than a straightforward, non-optimized implementation of the cosine II transform. Since H is the identity matrix, the non-negativity of x is guaranteed by the I-divergence constraint. Therefore, we can simplify the algorithm by omitting the constraint $x \geq 0$ and thus the variables y_3 and q_3 in the algorithm. This is equally true for the PDHGMp Algorithm II, where $s\,t < 1/\|(I\ \ L^{\mathrm{T}})\|_2^2$ is guaranteed for $s\,t < 1/9$. Alternatively, we can also use the predictor-corrector ADM Algorithm III, here. In Table 1 a speed comparison between these algorithms for 'trial and error' optimized parameters γ, s and t with respect to the error measure $\|x^{(k)} - x^*\|_\infty$ is provided. The sufficiently converged reference result x^* is obtained by Algorithm III. As the

PSNR: 22.63   PSNR: 27.03

Figure 4: *Left:* Original image of the French city of Nîmes (512×512) with values in the range $[1, 256]$, see also [32]. *Middle:* Image corrupted by multiplicative Gamma noise $(K = 10)$. *Right:* Restoration result by the *I*-divergence constrained model (16) with total variation seminorm $(\gamma = 0.015)$.

21

Algorithms	Parameters			Number of iterations	Computation times
	γ	s	t		
				$\|x^{(k)} - x^*\|_\infty < 3$	
Algorithm I: ADMM	0.042	–	–	36	2.3 sec
Algorithm I: ADMM with fixed $\lambda = 3.2286$	0.035	–	–	33	1.5 sec
Algorithm II: PDHGMp	–	4.6	0.033	44	1.4 sec
Algorithm III: ADM with predictor-corr. step	–	3.18	$\frac{1}{17}$	70	2.0 sec
"	–	(1)	$\left(\frac{1}{16}\right)$	(222)	(5.9 sec)
				$\|x^{(k)} - x^*\|_\infty < 1$	
Algorithm I: ADMM	0.055	–	–	66	4.6 sec
Algorithm I: ADMM with fixed $\lambda = 3.2286$	0.058	–	–	67	3.6 sec
Algorithm II: PDHGMp	–	3.5	0.043	81	2.2 sec
Algorithm III: ADM with predictor-corr. step	–	3.06	$\frac{1}{17}$	95	2.6 sec
"	–	(1)	$\left(\frac{1}{16}\right)$	(284)	(7.6 sec)

Table 1: Computation times required by the algorithms to compute $x^{(k)}$ with specified maximal pixel differences to the sufficiently converged reference result x^* of size 512×512 in Figure 4. The times are averaged here over 100 runs of the algorithms.

comparison shows Algorithm II is fastest here followed by Algorithm III if we optimize s and t *disregarding the theoretical convergence constraints* $s\,t < 1/\|(I\ L^{\mathrm{T}})\|_2^2$ and $s\,t < 0.5/\|L\|_2^2$, respectively. For the non-optimized values $s = 1/16$ and $t = 1$ used in [71] Algorithm III performs worse.

The ADMM Algorithm I is slightly slower than Algorithms II and III with optimized values s and t, here. However, this algorithm has the benefit that we only need to optimize one instead of two parameters and that convergence is theoretically assured for any $\gamma > 0$. Strategies for an adaptive parameter selection of γ for ADMM have been studied in [12, 43] and it is future work to adapt these methods for our algorithms. To get additionally a feeling about the performances compared to solving the penalized problem (17) we also executed Algorithm I with fixed, already optimized λ. In this case the algorithm is faster, but not significantly compared to the case where λ has to found by inner Newton iterations.

Nonlocal Regularization As mentioned in the introduction alternatively to the total variation seminorm, nonlocal terms $\|\,|Lx|\,\|_1$ can also be used in the restoration models. These methods often lead to better restoration results than TV-regularized approaches, but are computationally more demanding, since the matrix L is adapted to the image and is not as sparse as the discrete gradient matrix. For multiplicative Gamma noise appropriate nonlocal matrices L can be constructed as follows, compare [39, 63, 64]:

Remark 6.1. *We start with a zero weight matrix* $w \in \mathbb{R}^{n,n}$. *For every image pixel i we*

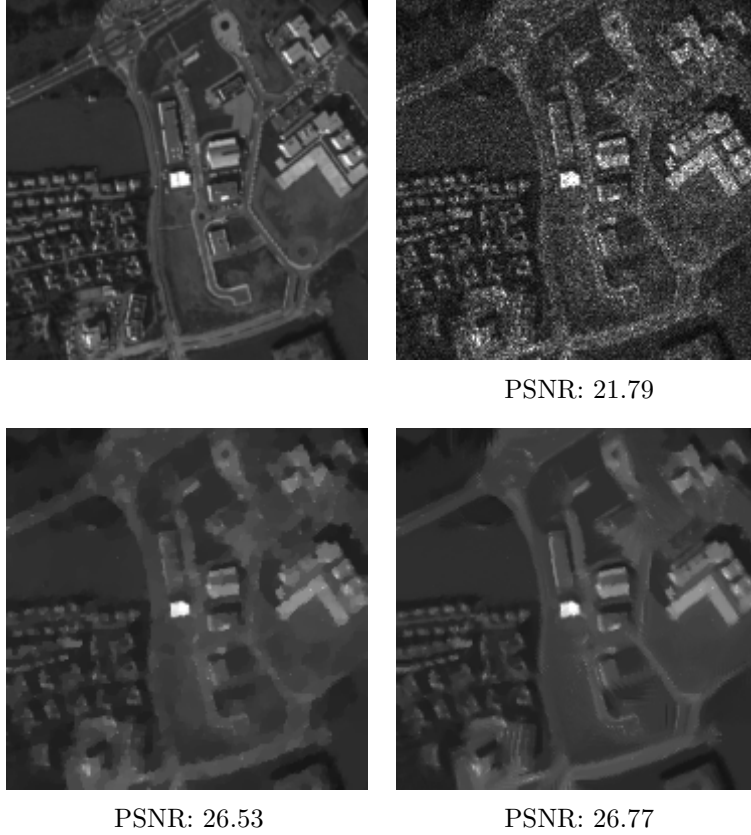PSNR: 21.79

PSNR: 26.53   PSNR: 26.77

Figure 5: *Top:* Parts of the images depicted in Fig. 4 (left and middle). *Bottom:* Restored images by the I-divergence constrained model (16) with total variation seminorm (left) and the nonlocal term ($p = 17$, $\omega = 17$, $a = 4$) (right), respectively.

compute for all j within a search window of size $\omega \times \omega$ around i the distances

$$d_a(i,j) := \sum_{h_1=-\lceil\frac{l-1}{2}\rceil}^{\lceil\frac{l-1}{2}\rceil} \sum_{h_2=-\lceil\frac{l-1}{2}\rceil}^{\lceil\frac{l-1}{2}\rceil} g_a(h_1, h_2)\, s\Big(f\big(i + (h_1, h_2)\big),\, f\big(j + (h_1, h_2)\big)\Big),$$

where $s(f_i, f_j) := K \log\left(\frac{2 + f_i/f_j + f_j/f_i}{4}\right)$ and g_a represents a discrete normalized Gaussian of mean 0 and standard deviation a. The parameter l controls here the size of the image parts being compared. For a predefined bound $\widetilde{m} = 5$ we select the $k \leq \widetilde{m}$ 'neighbors' $j \neq i$ of i for which $d_a(i,j)$ takes the smallest values and the number of nonzero elements in the row $w(j, \cdot)$ is smaller than $2\widetilde{m}$. Here, we set $w(i,j) = w(j,i) = 1$, which causes several weights $w(j, \cdot)$ to be already non-zero before we actually reach pixel j. To avoid that the number of non-zero weights becomes too large, we set the number of chosen neighbors to $k := \min\{\widetilde{m}, 2\widetilde{m} - r\}$ with r being the number of non-zero weights $w(i, \cdot)$ before the selection. Finally, we construct the matrix $L \in \mathbb{R}^{dn,n}$ with $d = 2\widetilde{m}$ so that L consists of d blocks of size $n \times n$, each having maybe some zero rows and rows with -1 as diagonal element plus one additional nonzero value 1 whose position is determined by the nonzero weights $w(i,j)$.

For such a matrix the constrained problem (16) with the estimated bound $\tau = 2.64\,n$ leads

Algorithms	Parameters			Number of iterations	Computation times
	γ	s	t		
				$\|x^{(k)} - x^*\|_\infty < 3$	
Algorithm I: ADMM	0.044	–	–	34	1.24 sec
Algorithm I: ADMM with fixed $\lambda = 6.4454$	0.029	–	–	23	0.73 sec
Algorithm II: PDHGMp	–	1.65	$\frac{1}{38}$	58	0.89 sec
Algorithm III: ADM with predictor-corr. step	–	4	$\frac{1}{80}$	30	1.04 sec
				$\|x^{(k)} - x^*\|_\infty < 1$	
Algorithm I: ADMM	0.047	–	–	46	1.46 sec
Algorithm I: ADMM with fixed $\lambda = 6.4454$	0.035	–	–	44	1.07 sec
Algorithm II: PDHGMp	–	1.65	$\frac{1}{38}$	76	1.22 sec
Algorithm III: ADM with predictor-corr. step	–	2.4	$\frac{1}{47}$	67	2.32 sec

Table 2: Computation times required by the algorithms to compute $x^{(k)}$ with specified maximal pixel differences to the sufficiently converged reference result x^* of size 180×180 shown in Figure 5 (right). The times are averaged here over 100 runs of the algorithms.

to even better restoration results than the total variation seminorm, see Figure 5.
Table 2 shows a time comparison of the algorithms for solving problem (16) with the nonlocal matrix ($d = 10$). The PDHGMp Algorithm II is here again slightly faster than the other algorithms. However, compared to approximately 7.2 seconds which we require for the construction of the matrix L, the time differences between the algorithms are almost negligible.

To finally conclude this section we also provide a result for a real synthetic aperture radar (SAR) image in Figure 6. Due to the image acquisition process the squared image values are corrupted by multiplicative Gamma noise ($K \approx 2.6$), here. Therefore, the I-divergence model (16) is applied to the squared image values and we take the square root of the output as final result. To compute the restoration result of (16) with the TV seminorm such that $\|x^{(k)} - x^*\|_\infty < 3$, ADMM Algorithm I needs approximately 5.3 seconds (97 iterations, $\gamma = 0.00013$) compared to 3 seconds required by PDHGMp Algorithm II (105 iterations, $s = 1150$, $t = \frac{1}{7400}$). To compute the result of (16) with the nonlocal matrix ($d = 10$) and the same accuracy, ADMM Algorithm I takes approximately 11.7 seconds (40 iterations, $\gamma = 0.00024$) compared to 11.2 seconds required by PDHGMp Algorithm II (99 iterations, $s = 550$, $t = \frac{1}{13000}$). Due to the image size the generation of the matrix L takes here with 48 seconds by far the most time.

7   Summary and Conclusions

We have examined the minimization of semi-norms under I-divergence constraints $D(H\cdot, b) \leq \tau$. One advantage of these models over penalized ones is the fact that the constraining parameter τ can be estimated by statistical methods if some knowledge on the (type of) noise is available. Primal-dual minimization algorithms for the constrained problem involve the

Figure 6: *Left:* Real multi-look SAR image (copyright by [34]) of size 512×512 with values in $[0, 255]$. *Middle:* Restoration result by the I-divergence constrained model (16) with the total variation seminorm applied to the squared image. *Right:* Result by the I-divergence constrained model (16) with the nonlocal term ($l = 9$, $\omega = 13$, $a = 3$) applied to the squared image.

solution of I-divergence constrained least squares problems in each iteration step k. To solve these problems one can utilize that the corresponding least squares problem with penalized I-divergence has an analytical solution $\hat{t}^{(k)}(\lambda^{(k)})$ which depends on the regularization parameter $\lambda^{(k)}$. Fortunately, there exists a unique parameter $\hat{\lambda}^{(k)}$ which solves the discrepancy equation $D(b, \hat{t}^{(k)}(\lambda^{(k)})) = \tau$. This $\hat{\lambda}^{(k)}$ can be efficiently computed by Newton's method. It turns out that the sequence $\{\hat{\lambda}^{(k)}\}_k$ converges to the regularization parameter $\hat{\lambda}$ for which the penalized problem (17) has the same solution.

Future directions of research may include the following:

- In the analysis of Lemma A3 which influences the following results of the paper we have used that $D(b, \cdot)$ is strictly convex for $b > 0$. Indeed this is a restriction for various image processing tasks. Using the agreement that $0 \log 0 := 0$ the definition of the I-divergence can be generalized to $b \geq 0$. It remains to prove under which assumptions on H and L our analysis still valids.

- The modification of our approach to spatially adapted regularization parameter selection, see [22, 29, 44], will be interesting. For this task, further estimates of appropriate parameters τ will be useful.

- The application of multiplicative iterative update rules for incorporating the non-negativity constraint, cf. [6, 28] should be examined.

- Moreover, the determination of the parameters inherent in the algorithms, i.e., γ and s, t is ongoing research.

- Recently, variance stabilizing transforms (VST) [31, 20] as the (generalized) Anscombe transform were applied instead of the I-divergence to restore Poissonian data, see, e.g., [20, 31]. However, the variance stabilizing property was not completely utilized since

all the models are penalized once where a parameter estimation is still necessary. The solution of the corresponding constrained problems was not addressed up to now.

- Finally, we are interested in analyzing multiple constraints. Here it may be useful to apply the primal-dual algorithms suggested in [19, 67] and epigraphical projections as in [5, 24].

A   Appendix

A.1   Auxiliary Lemmata

The first lemma ensures the existence of τ_0 in (15).

Lemma A.1. *Let $H \in \mathbb{R}^{n,n}$ with $\mathcal{K} \neq \emptyset$. Then $\operatorname{argmin}_{x \geq 0} D(b, Hx) \neq \emptyset$ holds true.*

Proof. Let $\tau_0 := \inf_{x \geq 0} D(b, Hx)$ and $x^{(n)} \geq 0$ be a sequence with $\lim_{n \to \infty} D(b, Hx^{(n)}) = \tau_0$. We have the unique decomposition $x^{(n)} = x_1^{(n)} + x_0^{(n)}$ with $x_1^{(n)} \in \mathcal{R}(H^*)$ and $x_0^{(n)} \in \mathcal{N}(H)$. Since $D(b, H\cdot)$ is lower level-bounded on $\mathcal{R}(H^*)$ and $\lim_{n \to \infty} D(b, Hx_1^{(n)}) = \tau_0$, the sequence $\{x_1^{(n)}\}$ is bounded. Thus, there exists a convergent subsequence $\{x_1^{(n_j)}\}$ with $\lim_{j \to \infty} x_1^{(n_j)} = \hat{x}_1 \in \mathcal{R}(H^*)$ and since $D(b, H\cdot)$ is continuous,

$$\lim_{j \to \infty} D(b, Hx_1^{(n_j)}) = D(b, H\hat{x}_1) = \tau_0. \tag{30}$$

We still have that $x^{(n_j)} = x_1^{(n_j)} + x_0^{(n_j)} \geq 0$ for some $x_0^{(n_j)} \in \mathcal{N}(H)$. By the following reasons there exists $\hat{x}_0 \in \mathcal{N}(H)$ such that $\hat{x} := \hat{x}_1 + \hat{x}_0 \geq 0$: Assume that this is not the case. Then, the affine space $\hat{x}_1 + \mathcal{N}(H)$ and the polyhedral cone $\mathbb{R}^n_{\geq 0}$ have an empty intersection. By [55, p. 175, Corollary 19.3.3] both sets can be strongly separated by a hyperplane, i.e., $\|\hat{x}_1 + v - z\| \geq \delta > 0$ for all $v \in \mathcal{N}(H)$ and all $z \geq 0$. Thus,

$$\delta \leq \|\hat{x}_1 - x_1^{(n_j)} + x_1^{(n_j)} + v - z\| \leq \|\hat{x}_1 - x_1^{(n_j)}\| + \|x_1^{(n_j)} + v - z\| \quad \forall v \in \mathcal{N}(H), \forall z \geq 0.$$

However, this is a contradiction, since the last summand becomes zero for $v = x_0^{(n_j)} \in \mathcal{N}(H)$ and some $z \geq 0$, and $\|\hat{x}_1 - x_1^{(n_j)}\|$ becomes arbitrary small for j large enough.

Finally, we conclude by (30) that there exists $\hat{x} \in \operatorname{argmin}_{x \geq 0} D(b, Hx)$. □

Next, we provide some useful lemmas which were applied in Section 3. The first lemma is a generalization of a lemma from [25].

Lemma A.2. *Let $\mathbb{R}^n$ be decomposed as orthogonal sums $\mathbb{R}^n = U_1 \oplus U_2$ and $\mathbb{R}^n = V_1 \oplus V_2$ of subspaces U_1, U_2 and V_1, V_2, where $U_2 \cap V_2 = \{0\}$. Let $F, G : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, lower semi-continuous functions with*

$$F(x) = F(x + u_2), \quad G(x) = G(x + v_2)$$

for all $x \in \mathbb{R}^n$, $u_2 \in U_2$ and $v_2 \in V_2$. Set $f := F|_{U_1}$ and $g := G|_{V_1}$ and assume that the level sets $\mathrm{lev}_\alpha f$, $\mathrm{lev}_\beta g$ are nonempty and bounded for some $\alpha, \beta \in \mathbb{R}$. Moreover, let $J : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semi-continuous function which is bounded from below. If $\mathrm{dom} F \cap \mathrm{dom} G \cap \mathrm{dom} J \neq \emptyset$, then $F + G + J$ attains its finite minimum.

Proof. Since f, g are proper, convex and lower semi-continuous and $\mathrm{lev}_\alpha(f)$, $\mathrm{lev}_\beta(g)$ are nonempty and bounded for some $\alpha, \beta \in \mathbb{R}$, we know that f and g are level-bounded, i.e., all their level sets are bounded, cf. [55, Cor. 8.7.1]. Moreover, by the lower semi-continuity of f and g all these level sets are compact. With the properness and again the lower semi-continuity of f and g we can further conclude that f and g are bounded from below. Without loss of generality we may therefore assume $f \geq 0$, $g \geq 0$, $J \geq 0$, which yields also that $F \geq 0$ and $G \geq 0$.

Now, we want to show that $F+G+J$ is level-bounded. Since $\mathrm{dom} F \cap \mathrm{dom} G \cap \mathrm{dom} J \neq \emptyset$, there exist $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma} \in \mathbb{R}$ with $\mathrm{lev}_{\tilde{\alpha}}(F) \cap \mathrm{lev}_{\tilde{\beta}}(G) \cap \mathrm{lev}_{\tilde{\gamma}}(J) \neq \emptyset$. Following the same arguments as in [25, Lemma 3.1 i)] we obtain by $U_2 \cap V_2 = \{0\}$ and the boundedness of $\mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(f)$ and $\mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(g)$ that $\mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(F) \cap \mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(G)$ is bounded. Since $F, G \geq 0$, the level set $\mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(F + G) \subseteq \mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(F) \cap \mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(G)$ is bounded as well and non-empty due to the fact that $\mathrm{lev}_{\tilde{\alpha}+\tilde{\beta}}(F+G) \supseteq \mathrm{lev}_{\tilde{\alpha}}(F) \cap \mathrm{lev}_{\tilde{\beta}}(G) \neq \emptyset$. Since $F + G$ is proper, convex and lower semi-continuous, this implies by [55, Cor. 8.7.1] that $F + G$ is level-bounded and with $J \geq 0$ we obtain that $F + G + J$ is level-bounded, too. Using now that $\mathrm{dom} F \cap \mathrm{dom} G \cap \mathrm{dom} J \neq \emptyset$ and that F, G and J are proper and lower semi-continuous, we know that $F + G + J$ is also proper and lower semi-continuous. Thus, it finally follows by [56, Thm. 1.9] that $F + G + J$ attains its finite minimum. $\square$

The next lemma is taken from [25].

Lemma A.3. *Let the Euclidean space $\mathbb{R}^n$ be decomposed into the direct sum $\mathbb{R}^n = U_1 \oplus U_2$ of two subspaces U_1, U_2 and let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function which is strictly convex on U_1 and which inheres the translation invariance $F(x) = F(x + u_2)$ for all $x \in \mathbb{R}^n$ and $u_2 \in U_2$. Furthermore, let $G : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be any convex function with $\mathrm{dom} F \cap \mathrm{dom} G \neq \emptyset$. Then all $\hat{x}, \tilde{x} \in \mathrm{argmin}_{x \in \mathbb{R}^n}\{F(x) + G(x)\}$ fulfill $\hat{x} - \tilde{x} \in U_2$ and $F(\hat{x}) = F(\tilde{x})$, $G(\hat{x}) = G(\tilde{x})$.*

A.2 Duality

Let $f_1 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $f_2 : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, lower semi-continuous functions and $A \in \mathbb{R}^{d,n}$. Then the primal problem

$$(P) \quad \min_{x \in \mathbb{R}^n} \{f_1(x) + f_2(Ax)\}$$

can be rewritten as

$$(P) \quad \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \{f_1(x) + f_2(y) \quad \text{subject to} \quad Ax = y\}.$$

Using the Lagrangian $L(x, y, p) = f_1(x) + f_2(y) + \langle p, Ax - y \rangle$ the primal and dual problems read

$$(P) \quad \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \max_{p \in \mathbb{R}^d} \{f_1(x) + f_2(y) + \langle p, Ax - y \rangle\},$$

$$(D) \quad \max_{p \in \mathbb{R}^d} \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^d} \{f_1(x) + f_2(y) + \langle p, Ax - y \rangle\}$$

and applying the definition of the conjugate function this becomes

$$(P) \quad \min_{x \in \mathbb{R}^n} \max_{p \in \mathbb{R}^d} \left\{ f_1(x) - f_2^*(p) + \langle p, Ax \rangle \right\},$$

$$(D) \quad \max_{p \in \mathbb{R}^d} \min_{x \in \mathbb{R}^n} \left\{ f_1(x) - f_2^*(p) + \langle p, Ax \rangle \right\}.$$

For the minimizers $\hat{p}$ of the dual problem we have that

$$
\begin{aligned}
\hat{p} \in \underset{p \in \mathbb{R}^d}{\operatorname{argmin}} &\left\{ f_2^*(p) - \min_{x \in \mathbb{R}^n} \left\{ f_1(x) + \langle p, Ax \rangle \right\} \right\} \\
= \underset{p \in \mathbb{R}^d}{\operatorname{argmin}} &\left\{ f_2^*(p) + \max_{x \in \mathbb{R}^n} \left\{ \langle -A^*p, x \rangle - f_1(x) \right\} \right\} \\
= \underset{p \in \mathbb{R}^d}{\operatorname{argmin}} &\{ f_2^*(p) + f_1^*(-A^*p) \}.
\end{aligned}
\tag{31}
$$

References

[1] K. J. Arrow, L. Hurwicz, and H. Uzawa. Studies in linear and non-linear programming. In *Stanford Mathematical Studies in the Social Sciences*, volume II. Stanford University Press, Stanford, 1958.

[2] S. Babacan, R. Molina, and A. Katsaggelos. Parameter estimation in TV image restoration using variational distribution approximation. *IEEE Transactions on Image Processing*, 17(3):326–339, 2008.

[3] J. M. Bardsley and J. Goldes. Regularization parameter selection methods for ill-posed Poisson maximum likelihood estimation. *Inverse Problems*, 25(9):095005, 2009.

[4] J. M. Bardsley and A. Luttman. Total variation-penalized Poisson likelihood estimation for ill-posed problems. *Advances in Computational Mathematics*, 31(1–3):35–59, 2009.

[5] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011.

[6] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[7] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. IOP Publishing, Bristol, 1998.

[8] M. Bertero, P. Boccacci, G. Talenti, R. Zanella, and L. Zanni. A discrepancy principle for Poisson noise. *Inverse Problems*, 25(4):105004, 2010.

[9] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Belmont, Massachusetts, 2003.

[10] J. M. Bioucas-Dias and M. A. T. Figueiredo. Multiplicative noise removal using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19(7):1720–1730, 2010.

[11] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*, volume 7 of *Springer Series in Operations Research*. Springer, Berlin, 2000.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[13] E. Bratsolis and M. Sigelle. A spatial regularization method preserving local photometry for Richardson-Lucy restoration. *Astronomy and Astrophysics*, 375(3):1120–1128, 2001.

[14] C. Brune, A. Sawatzky, and M. Burger. Primal and dual Bregman methods with application to optical nanoscopy. *International Journal of Computer Vision*, 92(2):211–229, 2011.

[15] C. B. Burckhardt. Speckle in ultrasound B-mode scans. *IEEE Transactions on Sonics and Ultrasonics*, 25(1):1–6, 1978.

[16] D. Calvetti and L. Reichel. Tikhonov regularization of large linear problems. *BIT Numerical Mathematics*, 43(2):263–283, 2003.

[17] M. Carlavan and L. Blanc-Feraud. Two constrained formulations for deblurring Poisson noisy data. In *Proceedings on the IEEE International Conference (ICIP)*, volume art no. 6116646, pages 689–692. 2011.

[18] M. Carlavan and L. Blanc-Feraud. Sparse Poisson noisy image deblurring. *IEEE Transactions on Image Processing*, 21(4):1834–1846, 2012.

[19] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[20] C. Chaux, L. Blanc-Féraud, and J. Zerubia. Wavelet-based restoration methods: Application in 3d confocal microscopy images. In *Proc. SPIE Conf. Wavelets*, page 67010E, San Diego, 2007.

[21] C. Chaux, J.-C. Pesquet, and N. Pustelnik. Nested iterative algorithms for convex constrained image recovery problems. *SIAM Journal on Imaging Science*, 2(2):730–762, 2009.

[22] D.-Q. Chen and L.-Z. Cheng. Spatially adapted regularization parameter selection based on the local discrepancy function for Poissonian image deblurring. *Inverse Problems*, 28(1):015004, 2012.

[23] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64(1–3):81–101, 1994.

[24] G. Cherchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu. A proximal approach for constrained cosparse modelling. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[25] R. Ciak, B. Shafei, and G. Steidl. Homogeneous penalizers and constraints in convex image restoration. *Preprint University of Kaiserslautern*, 2012.

[26] I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

[27] I. Daubechies, M. Fornasier, and I. Loris. Accelerated projected gradient method for linear inverse problems with sparsity constraints. *Journal of Fourier Analysis and Applications*, 14(5-6):764–792, 2008.

[28] A. R. De Pierro. Multiplicative iterative methods in computed tomography. In G. T. Herman, A. K. Louis, and F. Natterer, editors, *Mathematical Methods in Tomography*, volume 1497 of *Lecture Notes in Mathematics*, pages 167–186. Springer, 1991.

[29] Y. Dong, M. Hintermüller, and M. M. Rincon-Camacho. Automated regularization parameter selection in multi-scale total variation models for image restoration. *Journal of Mathematical Imaging and Vision*, 40(1):82–104, 2011.

[30] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *ICML '08 Proceedings of the 25th International Conference on Machine Learning*, ACM New York, 2008.

[31] F.-X. Dupé, J. Fadili, and J.-L. Starck. A proximal iteration for deconvolving Poisson noisy images using sparse representations. *IEEE Transactions on Image Processing*, 18(2):310–321, 2009.

[32] S. Durand, J. Fadili, and M. Nikolova. Multiplicative noise removal using L1 fidelity on frame coefficients. *Journal of Mathematical Imaging and Vision*, 36(3):201–226, 2010.

[33] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(3):293–318, 1992.

[34] European Space Agency (ESA). Multilook SAR image.

[35] J. Fadili and G. Peyré. Total variation projection with first order schemes. *IEEE Transactions on Image Processing*, 20(3):657–669, 2011.

[36] M. A. T. Figueiredo and J. M. Bioucas-Dias. Restoration of Poissonian images using alternating direction optimization. *IEEE Transactions on Image Processing*, 19(12):3133–3145, 2010.

[37] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary–Value Problems*, volume 15 of *Studies in Mathematics and its Applications*, chapter 9, pages 299–331. Elsevier Science Publishers B.V., Amsterdam, 1983.

[38] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computational and Applied Mathematics*, 2(1):17–40, 1976.

[39] G. Gilboa and S. Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2008.

[40] I. S. Gradshteyn and I. Ryzhik. *Table of Integrals, Series, and Products*. Elsevier, 2007.

[41] M. Hanke and P. C. Hansen. Regularization methods for large-scale problems. *Survey on Mathematics for Industry*, 3:253–315, 1993.

[42] P. C. Hansen, M. E. Kilmer, and R. H. Kjeldsen. Exploiting residual information in the parameter choice for discrete ill-posed problems. *BIT Numerical Mathematics*, 46(1):41–59, 2006.

[43] B. S. He, H. Yang, and S. L. Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and Applications*, 106(2):337–356, 2000.

[44] M. Hintermüller and M. M. Rincon-Camacho. Expected absolute value estimators for a spatially adapted regularization parameter choice rule in L^1-TV-based image restoration. *Inverse Problems*, 26(8):085005, 2010.

[45] J.-B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms*, volume 1. Springer, Berlin, Heidelberg, 1993.

[46] K. V. Ivanov, V. V. Vasin, and V. P. Tanana. *Theory of Linear Ill-Posed Problems and its Applications*. Brill Academic Publishers, Utrecht, Boston, Koeln, Tokyo, 2002.

[47] E. Jonsson, S.-C. Huang, and T. Chan. Total variation regularization in positron emission tomography. CAM-Report 98-48, UCLA, Los Angeles, 1998.

[48] C. L. Lawson and R. J. Hansen. *Solving least squares problems*. Prentice-Hall, Englewood Cliffs, 1974.

[49] T. Le, R. Chartrand, and T. J. Asaki. A variational approach to reconstructing images corrupted by Poisson noise. *Journal of Mathematical Imaging and Vision*, 27(3):257–263, 2007.

[50] H. Maître. *Processing of Synthetic Aperture Radar Images*. ISTE Ltd and John Wiley & Sons, 2008.

[51] C. B. Morrey. *Multiple Integrals in the Calculus of Variations*. Springer, Berlin, 1966.

[52] M. K. Ng, P. Weiss, and X. Yuan. Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods. *SIAM Journal on Scientific Computing*, 32(5):2710–2736, 2010.

[53] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. T. Figueiredo. Adaptive total variation image deblurring: A majorization–minimization approach. *Signal Processing*, 89(9):1683–1693, 2009.

[54] T. Pock, A. Chambolle, D. Cremers, and H. Bischof. A convex relaxation approach for computing minimal partitions. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 810–817, 2009.

[55] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[56] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *A Series of Comprehensive Studies in Mathematics*. Springer, Berlin, 2 edition, 2004.

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

[58] B. W. Rust and D. P. O'Leary. Residual periodograms for choosing regularization parameters for ill-posed problems. *Inverse Problems*, 24(3):034005, 2008.

[59] I. W. Selesnik, R. G. Baraniuk, and N. C. Kingsbury. The dual-tree complex wavelet transform. *IEEE Transactions on Signal Processing*, 22(6):123–151, 2005.

[60] S. Setzer. Operator splittings, Bregman methods and frame shrinkage in image processing. *International Journal of Computer Vision*, 92(3):265–280, 2011.

[61] S. Setzer, G. Steidl, and T. Teuber. Deblurring Poissonian images by split Bregman techniques. *Journal of Visual Communication and Image Representation*, 21(3):193–199, 2010.

[62] S. Setzer, G. Steidl, and T. Teuber. On vector and matrix median computation. *Journal of Computational and Applied Mathematics*, 236(8):2200–2222, 2012.

[63] G. Steidl and T. Teuber. Removing multiplicative noise by Douglas-Rachford splitting methods. *Journal of Mathematical Imaging and Vision*, 36(2):168–184, 2010.

[64] T. Teuber and A. Lang. A new similarity measure for nonlocal filtering in the presence of multiplicative noise. *Computational Statistics & Data Analysis*, 56(12):3821–3842, 2012.

[65] T. Teuber, G. Steidl, and R. H. Chan. Minimization and parameter estimation for seminorm regularization models with I-divergence constraints. Technical report, University of Kaiserslautern, Kaiserslautern, Germany, 2012.

[66] E. van den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

[67] B. C. Vu. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, page accepted, 2012.

[68] R. F. Wagner, S. W. Smith, J. M. Sandrik, and H. Lopez. Statistics of speckle in ultrasound B-scans. *IEEE Transactions on Sonics and Ultrasonics*, 30(3):156–163, 1983.

[69] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, 14(4):651–667, 1977.

[70] P. Weiss, L. Blanc-Féraud, and G. Aubert. Efficient schemes for total variation minimization under constraints in image processing. *SIAM Journal on Scientific Computing*, 31(3):2047–2080, 2009.

[71] Y.-W. Wen and R. H. Chan. Parameter selection for total variation based image restoration using discrepancy principle. *IEEE Transactions on Image Processing*, 21(4):1770–1781, 2012.

[72] H. Woo and Y. Yun. Alternating minimization algorithm for speckle reduction with a shifting technique. *IEEE Transactions on Image Processing*, 21(4):1701–1714, 2012.

[73] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

[74] R. Zanella, P. Boccacci, L. Zanni, and M. Bertero. Efficient gradient projection methods for edge-preserving removal of Poisson noise. *Inverse Problems*, 25(4):045010, 2009.

[75] M. Zhu and T. F. Chan. An efficient primal-dual hybrid gradient algorithm for total vatiation image restoration. CAM-Report 08-34, UCLA, Los Angeles, 2008.