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

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

July 23, 2012

Abstract

This papers deals with the minimization of seminorms $\|L \cdot\|$ on $\mathbb{R}^n$ under the constraint of a bounded I-divergence $D(b, H\cdot)$. 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. Typically, H represents here, e.g., a linear blur operator and L is some discrete derivative operator. Our preference for the constrained approach over the corresponding penalized version is based on the fact that the I-divergence of data corrupted, e.g., by Poisson noise or multiplicative Gamma noise can be estimated by statistical methods. Our minimization technique rests upon relations between constrained and penalized convex problems and resembles the idea of Morozov's discrepancy principle. More precisely, we propose first-order primal-dual algorithms which reduce the problem to the solution of certain proximal minimization problems in each iteration step. The most interesting of these proximal minimization problems is an I-divergence constrained least squares problem. We solve this problem by connecting it to the corresponding I-divergence penalized least squares problem with an appropriately chosen regularization parameter. Therefore, our 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 penalized problem has the same solution as our constrained one. In other words, the solution of this penalized problem fulfills the I-divergence constraint. We provide the proofs which are necessary to understand our approach and demonstrate the performance of our algorithms for different image restoration examples.

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 [39] and the survey paper [34]. One of the best examined models in $\mathbb{R}^n$ is

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

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 not invertible or ill-conditioned. The linear operator $L \in \mathbb{R}^{m,n}$

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

in the regularization term enforces some regularity of the minimizer. Examples are discrete derivative operators or nonlocal operators as considered in [31, 55]. 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 topic, e.g., Morozov's discrepancy principle [44], the L-curve criterion [41], the generalized cross-validation [59], normalized cumulative or residual periodogram approaches [35, 51] and variational Bayes' approaches [2, 46]. In this paper, we will adapt the simple idea of the discrepancy principle, which chooses the regularization parameter such that the norm of the defect $\|b - Hx\|_2$ equals some known error.
When dealing with image processing applications the above model is often replaced by

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

with 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 [50] 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. Recently, also the constrained model

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

was handled in [45], which has the advantage that some knowledge on the noise allows to estimate the parameter τ rather than the regularization parameter λ. Similarly, in [15], the authors consider the problem from the point of view of the penalized problem (1). They propose a primal-dual algorithm with a predictor-corrector scheme [18] which resembles in some way the method in [30]. 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 $\hat{x}$ of the corresponding penalized problem which is also computed by the algorithm fulfills $\|H\hat{x} - b\|_2 \le \tau$. However, independently of the point of view of the authors, a common ingredient of all these algorithms is the fact that the solution $\hat{t}(\lambda)$ of the least squares problem with penalized $\|H\cdot - b\|_2^2$ term

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

is given analytically and that moreover $f(\lambda) := \|H\hat{t}(\lambda) - b\|_2^2 = \tau^2$ has a unique solution λ, which can be computed by certain methods, see [15, 32].
Note that other constrained models and corresponding efficient algorithms were recently proposed for image processing and sparsity promoting tasks, see e.g., [19, 21, 27, 57, 60].
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, cf. [3, 40, 42, 55, 62]. Poisson data typically occurs in imaging processes where the images are obtained by counting particles, e.g., photons, that hit the image domain, see [6]. Multiplicative noise often appears as speckle in applications like laser, ultrasonic [14, 58] or synthetic aperture radar (SAR) imaging [12, 43]. In the following, we want to solve the I-divergence constrained problem

$$\operatorname*{argmin}_{x\ge 0} \left\{ \|Lx\| \quad \text{subject to} \quad D(b, Hx) \le \tau \right\}, \tag{3}$$

where we also have to cope with a non-negativity constraint which is inherent in most applications. As for the above least squares - TV problems we propose primal-dual algorithms. Again these algorithms will relate the constrained problem to the penalized one

$$\operatorname*{argmin}_{x\geq 0} \{\|Lx\| + \lambda D(b, Hx)\} \tag{4}$$

with an appropriate regularization parameter via the discrepancy principle. Note that the penalized I-divergence - TV problem was also approached by Bregman-EM-TV methods [13]. Our primal-dual algorithms restrict the problem to the iterative solution of certain proximal minimization problems, see [4]. All these simpler proximal minimization problems can be solved by meanwhile standard methods except the I-divergence least squares problem

$$\operatorname*{argmin}_{t\in\mathbb{R}^n} \left\{ \frac{1}{2}\|t-a\|_2^2 \quad \text{subject to} \quad D(b,t) \leq \tau \right\}. \tag{5}$$

Here, we use that there exists an analytical expression for the minimizer $\hat{t}(\lambda)$ of the penalized least squares problem

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

and that moreover,

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

has a unique solution λ which can be computed, e.g., by Newton's method. Once we have found this λ, we can compute $\hat{t}(\lambda)$ using the analytical expression. Of course, this $\hat{t}(\lambda)$ solves our constrained problem (5). As end product our algorithm computes the minimizer $\hat{x}$ of (3) and as a by-product the regularization parameter $\hat{\lambda}$ such that $\hat{x}$ is also a solution of the penalized problem (4) with this regularization parameter.

The structure of this paper is as follows: In Section 2 we provide the basic notation and a theorem on the general relation between constrained and penalized convex problems. Since an important step of our minimization algorithms consists in the solution of least squares problems with constrained I-divergence we study these problems in Section 3. Section 4 analyzes the penalized problem (4) and the constrained problem (3). We will see that under mild assumptions both problems have solutions and that different solutions of the same problem leave $\|L\cdot\|$ and $H\cdot$ fixed. Moreover, we clarify the relation between the constrained and the penalized problem, which is central for understanding the convergence properties of the subsequent algorithms. In Section 5, we deal with the minimization of the constrained problem by primal-dual algorithms. First, we introduce the dual problems and consider their relations to the primal ones. Then, we apply an ADMM algorithm together with an algorithm proposed in Section 3 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 (3) and on the other hand computes the regularization parameter $\hat{\lambda}$ such that the penalized problem (4) has the same solution. Next, we discuss the application of other primal-dual algorithms. The main ingredient of all these algorithms is again the solution of inner least squares problems with I-divergence constraints. In Section 6, we show how to determine appropriate choices for the parameter τ in the cases of Poisson noise and multiplicative Gamma-distributed noise. In contrast to the regularization parameter λ in (4) a reasonable value for τ in (3) can usually

be directly determined by statistical considerations if the type of noise corrupting the data is known. Section 7 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 8. The Appendix contains some auxiliary lemmas and provides standard relations on dual problems.

2 Notation and Basic Relations

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 and by

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

the *indicator function* ι_S of a set $S \neq \emptyset$. For a norm we have

$$\partial\|x\| = \begin{cases} B_{\|\cdot\|_*}(1) & \text{if } \|x\| = 0, \\ \{p \in \mathbb{R}^n : \langle p, x \rangle = \|x\|, \|p\|_* = 1\} & \text{otherwise} \end{cases} \tag{6}$$

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 $S := \mathbb{R}^n_{\geq 0}$ and $x \geq 0$, we have for example

$$\partial\iota_{\mathbb{R}^n_{\geq 0}}(x) = N_{\mathbb{R}^n_{\geq 0}}(x) = \mathcal{I}_1 \times \ldots \times \mathcal{I}_n, \quad \text{where} \quad \mathcal{I}_k := \begin{cases} (-\infty, 0] & \text{if } x_k = 0, \\ \{0\} & \text{if } x_k > 0 \end{cases} \tag{7}$$

and $\iota^*_{\mathbb{R}^n_{\geq 0}} = \sigma_{\mathbb{R}^n_{\geq 0}} = \iota_{\mathbb{R}^n_{\leq 0}}$.
For given $b \in \mathbb{R}^n_{>0}$ 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. [20]. Note that

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

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

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

Finally, it will be useful to notice the following well-known relation between constrained and penalized convex problems, see, e.g., [38, 19].

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*

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \{G(x) + \lambda F(x)\}, \qquad \lambda \geq 0 \tag{8}$$

and

$$\operatorname*{argmin}_{x \in \mathbb{R}^n} \{G(x) \quad \text{subject to} \quad F(x) \leq \tau\} \tag{9}$$

are related as follows:
i) *Assume that* $\mathrm{dom}\, F \cap \mathrm{dom}\, G \neq \emptyset$. *Let $\hat{x}$ be a minimizer of* (8). *If $\lambda = 0$, then $\hat{x}$ is also a minimizer of* (9) *if and only if $\tau \geq F(\hat{x})$. If $\lambda > 0$, then $\hat{x}$ is also a minimizer of* (9) *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* (9). *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* (8). *If $\hat{x}$ is in addition not a minimizer of G, then $\lambda > 0$.*

Concerning i) we mention that in case the minimizer of (8) 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 [19].

We will apply Theorem 2.1 with respect to the functions $F := D(b, H\cdot)$ and $G := \|L\cdot\| + \iota_{\mathbb{R}^n_{\geq 0}}$. Then we will see that for appropriately chosen τ and a solution $\hat{x}$ of (9) there exists a unique λ such that $\hat{x}$ is also a solution of (8).

3 Least Squares - I-Divergence Problems

The main part of our algorithms for solving (3) will consist in the solution of least squares problems with constrained I-divergence, i.e., in solving problems of the form

$$\operatorname*{argmin}_{t \in \mathbb{R}^n} \{\frac{1}{2}\|t - a\|_2^2 \quad \text{subject to} \quad D(b,t) \leq \tau\}, \quad \tau \geq 0 \tag{10}$$

with $a \in \mathbb{R}^n$. Therefore, we deal with these simpler problems first. We will solve these constrained least squares problems by utilizing the corresponding penalized problems

$$\operatorname*{argmin}_{t \in \mathbb{R}^n} \{\frac{1}{2}\|t - a\|_2^2 + \lambda D(b,t)\}, \quad \lambda \geq 0. \tag{11}$$

Both problems have a solution which is moreover unique, since the functionals are coercive and strictly convex. 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 theorem.

Theorem 3.1. *Let $a, b \in \mathbb{R}^n$ with $b > 0$ and $a \neq b$ be given.*
i) *Let $\lambda = 0$. Then problem* (11) *has the solution $\hat{t} = a$. This is also a solution of* (10) *if and only if $a > 0$ and $\tau \geq D(b, a)$. For $\lambda > 0$ problem* (11) *has the solution*

$$\hat{t} \;=\; g(a, \lambda) \;=\; \frac{1}{2}\left(a - \lambda + \sqrt{(a-\lambda)^2 + 4\lambda b}\right), \tag{12}$$

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

$$f(\lambda) \;:=\; D(b, g(a,\lambda)) \;=\; \langle 1_n, g(a,\lambda)\rangle - \langle b, \log g(a,\lambda)\rangle - \langle 1_n, b - b\log b)\rangle$$

is strictly decreasing and $\hat{t}$ is also the solution of (10) *exactly for $\tau = D(b, g(a, \lambda))$.*
ii) *Let $\tau = 0$. Then problem* (10) *has the solution $\hat{t} = b$ and there does not exist $\lambda \geq 0$ such that $\hat{t} = b$ is the solution of* (11)*. Let $\tau > 0$. Then the unique solution $\hat{t} > 0$ of problem* (10) *has the following properties: If $a > 0$ and $D(b, a) \leq \tau$, then $\hat{t} = a$ and this is also the solution of* (11) *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* (11)*.*

Proof. i) Let $\lambda = 0$. Then problem (11) has obviously the solution $\hat{t} = a$, which can only be a solution of (10) if and only if $a > 0$ and $D(b, a) \leq \tau$.
Let $\lambda > 0$. Then the minimizer of (11) can be computed separately for each component with index $i = 1, \ldots, n$. Setting the gradient of

$$\frac{1}{2}(a_i - t_i)^2 + \lambda D(b_i, t_i) = \frac{1}{2}(a_i - t_i)^2 + \lambda(t_i - b_i \log t_i) - b_i + b_i \log b_i$$

to zero, we obtain the quadratic equation

$$t_i^2 - t_i(a_i - \lambda) - b_i\lambda = 0,$$

which has the positive solution

$$\hat{t}_i = \frac{1}{2}\left(a_i - \lambda + \sqrt{(a_i - \lambda)^2 + 4\lambda b_i}\right).$$

This proves (12). Since $a \neq b$, $b > 0$ and $\lambda > 0$, we see that $\hat{t} \notin \{a, b\}$.
Let $\hat{t} = \hat{t}(\lambda) = g(a, \lambda)$. Next, we prove that $f(\lambda) = D(b, \hat{t}(\lambda))$ is strictly decreasing. Since f is up to a constant the sum of the functions $\hat{t}_i(\lambda) - b_i \log \hat{t}_i(\lambda)$, $i = 1, \ldots, n$, it is sufficient to show that the monotonicity relation holds true in one dimension, i.e., for $n = 1$. Thus, it remains to prove that

$$f'(\lambda) = \hat{t}'(\lambda) - \hat{t}'(\lambda)\frac{b}{\hat{t}(\lambda)} < 0,$$

where

$$\hat{t}'(\lambda) = \frac{1}{2}\left(-1 + \frac{\lambda - a + 2b}{\sqrt{(a-\lambda)^2 + 4\lambda b}}\right) = \frac{-\hat{t}(\lambda) + b}{\sqrt{(a-\lambda)^2 + 4\lambda b}}. \tag{13}$$

We see that $\hat{t}'(\lambda) = 0$ if and only if $a = b$, which is not possible by assumption. Further, we have

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

Finally, we obtain the rest of the assertion i) by applying Theorem 2.1i).
ii) Let $\tau = 0$. Then, problem (10) has obviously the solution $\hat{t} = b$. By part i) it follows immediately that $\hat{t} = b$ is not a solution of (11) for any $\lambda \geq 0$.
Let $\tau > 0$ and let $\hat{t} > 0$ be the unique solution of (10). If $a > 0$ and $D(b, a) \leq \tau$, then obviously $\hat{t} = a$ and this is also the solution of (11) for $\lambda = 0$. By i) we see that $\hat{t} = a$ cannot be a solution of (11) for $\lambda > 0$. If a is not componentwise positive or $\tau < D(b, a)$, then $\hat{t} = a$ is not the solution of (10). Moreover, $\hat{t} = b$ is also not the solution of (10) by the following argument: Since $\tau > 0$ and $D(b, \cdot)$ is continuous there exists a neighborhood of b such that $D(b, t) \leq \tau$ for all t in this neighborhood, in particular for $t = b - \mu(b - a)$ and μ small enough. Then $\frac{1}{2}\|b - \mu(b-a) - a\|_2^2 = \frac{(1-\mu)^2}{2}\|a - b\|_2^2$ is smaller than $\frac{1}{2}\|a - b\|_2^2$ for $\mu > 0$.
Now we can apply Theorem 2.1 ii) and conclude that there exists a unique $\lambda > 0$ such that $\hat{t}$ is also a solution of (11). This completes the proof. □

Note that it was proved in [7] 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. However, Theorem 3.1 describes $D(b, \hat{t}(\lambda))$ more detailed in our special case.
Based on Theorem 3.1 we can use the following algorithm to find the solution $\hat{t}$ of the least squares problem with I-divergence constraint (10):

Algorithm I (Solution of (10)**)**
Input: $a, b \in \mathbb{R}^n$, $b > 0$ and $\tau > 0$.
Find $\hat{\lambda}$ as the unique solution of

$$f(\lambda) = \tau$$

by Newton's method. Set

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

Using (13) and (14) we obtain for the derivative required in the Newton method

$$f'(\lambda) = -\sum_{i=1}^{n} \frac{(-g_i(a_i, \lambda) + b_i)^2}{g_i(a_i, \lambda)\sqrt{(a_i - \lambda)^2 + 4\lambda b_i}}.$$

4 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 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 4.1. *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* (30) *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*

$$\begin{aligned} D(b, \alpha h) &= \langle 1_n, \alpha h - b\log(\alpha h)\rangle - \langle 1_n, b - b\log b\rangle \\ &= \alpha n - \langle 1_n, b\log(\alpha h)\rangle - \langle 1_n, b - b\log b\rangle \end{aligned}$$

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

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

and

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

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*

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

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 4.2. *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* $(P_{1,\tau})$ *and* $(P_{2,\lambda})$ *have a solution.*

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

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

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* $D(b, H\hat{x}) = \tau$.

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 $G := \|L\cdot\|$, $g := G|_{\mathcal{R}(L^*)}$, $J := \iota_{\mathbb{R}^n_{\geq 0}}$ 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 $F := D(b, H\cdot)$, $G := \|L\cdot\| + \iota_{\mathbb{R}^n_{\geq 0}}$ 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 [8, Proposition 4.7.2] that since $\mathrm{lev}_\tau D(b, H\cdot) \cap \mathbb{R}^n_{\geq 0} \neq \emptyset$ and $\|L\cdot\|$ is continuous on its domain $\mathbb{R}^n$, there exists $v \in \partial\|L\cdot\|(\hat{x})$ such that

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

We have that $v = L^*\partial\|L\hat{x}\|$. Since $\tau < \tau_L$, we know that $\hat{x} \notin \mathcal{N}(L)$. Thus, by (6), $v = L^*\hat{p}$ 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. □

Lemma 4.3. *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 $\mathrm{argmin}_{x\geq 0} D(b, Hx) \cap \mathcal{N}(L) = \emptyset$. Let $\hat{x}$ be a solution of $(P_{2,\lambda})$ 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 $\hat{x} \in \mathrm{argmin}_{x\geq 0}\{\|Lx\| + \lambda D(b, Hx)\}$ if and only if $\hat{x} \geq 0$ and

$$\begin{aligned}
0 &\in \partial\left(\|L\cdot\| + \lambda D(b, H\cdot) + \iota_{\mathbb{R}^n_{\geq 0}}\right)(\hat{x}),\\
0 &\in L^*\partial\|L\hat{x}\| + \lambda H^*\nabla D(b, H\hat{x}) + \partial\iota_{\mathbb{R}^n_{\geq 0}}(\hat{x}),\\
0 &\in L^*\partial\|L\hat{x}\| + \lambda H^*\left(1_n - \frac{b}{H\hat{x}}\right) + N_{\mathbb{R}^n_{\geq 0}}(\hat{x}),\\
\lambda H^*\left(\frac{b}{H\hat{x}} - 1_n\right) &\in L^*\partial\|L\hat{x}\| + N_{\mathbb{R}^n_{\geq 0}}(\hat{x}).
\end{aligned}$$

By (6) 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 (7) 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 $D(b, H\hat{x}) = \tau_L$. □

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

Theorem 4.4. *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 $(P_{1,\tau})$ with $\tau_0 < \tau < \tau_L$, then there exists a unique $\lambda > 0$ such that $\hat{x}$ is also a solution of $(P_{2,\lambda})$. 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 4.3 this λ is uniquely determined and by Theorem 4.2iii) it does not depend on the chosen solution of $(P_{1,\tau})$. □

5 Minimization of Seminorms with Constrained I-Divergence

In this section, we compute a solution of $(P_{1,\tau})$ for $\tau_0 < \tau < \tau_L$. First, we will apply an ADMM algorithm together with Algorithm I 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. The main ingredient of all these algorithms is again the solution of inner least squares problems with I-divergence constraints.
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.

5.1 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

$$(P_{1,\tau}) \quad \underset{\substack{x\in\mathbb{R}^n \\ y\in\mathbb{R}^{2n+m}}}{\operatorname{argmin}} \Big\{ \underbrace{\langle 0, x\rangle}_{=:f_1(x)} + \underbrace{\iota_{\operatorname{lev}_\tau D(b,\cdot)}(y_1) + \|y_2\| + \iota_{y_3\geq 0}(y_3)}_{:=f_2(y_1,y_2,y_3)} \text{ s.t. } \underbrace{\begin{pmatrix} H \\ L \\ I \end{pmatrix}}_{A} x = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \Big\}, \quad (21)$$

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

Using the duality relations in Appendix A.2, in particular (32), 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

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

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

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 $\mathrm{SOL}(X)$ denote the solution set of problem (X).

Lemma 5.1. *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:*

$$\left.\begin{array}{l} \hat{x} \in \mathrm{SOL}(P_{1,\tau}) \\ \hat{p} \in \mathrm{SOL}(D_{1,\tau}) \end{array}\right\} \quad \Leftrightarrow \quad \left\{\begin{array}{c} \hat{p}_1 \in N_{\operatorname{lev}_\tau D(b,\cdot)}(H\hat{x}),\ \hat{p}_2 \in \partial\|L\hat{x}\|,\ \hat{p}_3 \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x}) \\ \textit{such that} \quad H^*\hat{p}_1 + L^*\hat{p}_2 + \hat{p}_3 = 0, \end{array}\right. \quad (22)$$

and

$$\left.\begin{array}{l} \hat{x} \in \mathrm{SOL}(P_{2,\lambda}) \\ \hat{p} \in \mathrm{SOL}(D_{2,\lambda}) \end{array}\right\} \quad \Leftrightarrow \quad \left\{\begin{array}{c} \hat{p}_1 = \lambda(1_n - \frac{b}{H\hat{x}}),\ \hat{p}_2 \in \partial\|L\hat{x}\|,\ \hat{p}_3 \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x}) \\ \textit{such that} \quad H^*\hat{p}_1 + L^*\hat{p}_2 + \hat{p}_3 = 0. \end{array}\right. \quad (23)$$

Since $\mathrm{SOL}(P_{1,\tau})$ and $\mathrm{SOL}(P_{2,\lambda})$ are nonempty, the proof follows by standard arguments from the duality theory of convex functions, cf. [10].
The following subsections describe algorithms to solve $(P_{1,\tau})$.

5.2 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 [53], see also [9, 28]. Considering $(P_{1,\tau})$ in the form (21) we obtain the following algorithm:

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

$$\begin{aligned}
x^{(k+1)} &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \big\{\|q_1^{(k)} + Hx - y_1^{(k)}\|_2^2 + \|q_2^{(k)} + Lx - y_2^{(k)}\|_2^2 + \|q_3^{(k)} + x - y_3^{(k)}\|_2^2\big\},\\
y_1^{(k+1)} &= \operatorname*{argmin}_{y_1\in\mathbb{R}^n} \big\{\iota_{\mathrm{lev}_\tau D(b,\cdot)}(y_1) + \frac{\gamma}{2}\|q_1^{(k)} + Hx^{(k+1)} - y_1\|_2^2\big\},\\
y_2^{(k+1)} &= \operatorname*{argmin}_{y_2\in\mathbb{R}^m} \big\{\|y_2\| + \frac{\gamma}{2}\|q_2^{(k)} + Lx^{(k+1)} - y_2\|_2^2\big\},\\
y_3^{(k+1)} &= \operatorname*{argmin}_{y_3\in\mathbb{R}^n} \big\{\iota_{y_3\geq 0}(y_3) + \frac{\gamma}{2}\|q_3^{(k)} + x^{(k+1)} - y_3\|_2^2\big\},\\
q_1^{(k+1)} &= q_1^{(k)} + Hx^{(k+1)} - y_1^{(k+1)},\\
q_2^{(k+1)} &= q_2^{(k)} + Lx^{(k+1)} - y_2^{(k+1)},\\
q_3^{(k+1)} &= q_3^{(k)} + x^{(k+1)} - y_3^{(k+1)}.
\end{aligned} \tag{24}$$

Note that this is a so-called *scaled ADMM algorithm* where $q = p/\gamma$ replaces the dual variable p. The above minimization problems are strictly convex problems, which have a unique solution.
The first minimization problem in the algorithm is a least squares problem whose unique solution is given by the solution of a linear system of equations:

$$x^{(k+1)} = (H^{\mathrm{T}}H + L^{\mathrm{T}}L + I)^{-1}\big(H^{\mathrm{T}}(y_1^{(k)} - q_1^{(k)}) + L^{\mathrm{T}}(y_2^{(k)} - q_2^{(k)}) + (y_3^{(k)} - q_3^{(k)})\big). \tag{25}$$

This linear system of equations can often be efficiently solved by a conjugate gradient (CG) method. Sometimes, when the orthogonal decomposition of $H^{\mathrm{T}}H + L^{\mathrm{T}}L + I$ is known, it is even possible to solve it explicitly. This is in particular the case for Gaussian blur matrices H and the discrete gradient $L = \nabla$ with reflecting boundary conditions. In this case, the matrix $H^{\mathrm{T}}H + L^{\mathrm{T}}L + I$ can be diagonalized by the discrete cosine II transform.
Denoting by P_C the *orthogonal projection* onto a set C we obtain further

$$\begin{aligned}
y_2^{(k+1)} &= \big(I - P_{B_{\|\cdot\|_*}(1/\gamma)}\big)\big(q_2^{(k)} + Lx^{(k+1)}\big),\\
y_3^{(k+1)} &= P_{\mathbb{R}_{\geq 0}}\big(q_3^{(k)} + x^{(k+1)}\big).
\end{aligned}$$

The orthogonal projection onto $B_{\|\cdot\|_*}(1/\gamma)$ can be easily computed for the ℓ_p-norms with $p = 1, \infty$ and their mixed versions, see, e.g., [24, 54, 63].
The interesting part is the computation of $y_1^{(k+1)}$ in (24). Setting $a^{(k+1)} := q_1^{(k)} + Hx^{(k+1)}$ we see that

$$y_1^{(k+1)} = \operatorname*{argmin}_{t\in\mathbb{R}^n}\{\frac{\gamma}{2}\|t - a^{(k+1)}\|_2^2 \quad \text{subject to} \quad D(b,t) \leq \tau\}.$$

By Theorem 3.1ii) we have the following: If $a^{(k+1)} > 0$ and $D(b, a^{(k+1)}) \leq \tau$, then

$$y_1^{(k+1)} = a^{(k+1)} \quad \text{and} \quad \lambda_{k+1} = 0.$$

Otherwise, there exists a unique $\lambda_{k+1} > 0$ such that

$$y_1^{(k+1)} = \operatorname*{argmin}_{t\in\mathbb{R}^n}\{\frac{\gamma}{2}\|t - a^{(k+1)}\|_2^2 + \lambda_{k+1}D(b,t)\}.$$

By Theorem 3.1i) we obtain that

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

where λ_{k+1} is the unique solution of

$$f(\lambda) = D\left(b, g\left(a^{(k+1)}, \frac{\lambda}{\gamma}\right)\right) = \tau$$

and $g : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}^n$ is defined by (12). This solution can be computed, e.g., by Newton's method with initial value λ_k. In summary we have:

Algorithm II (ADMM with inner Newton iterations)
Initialization: $q_1^{(0)} = q_2^{(0)} = q_3^{(0)} = 0$, $y_1^{(0)} = Hb$, $y_2^{(0)} = Lb$, $y_3^{(0)} = b$, $\lambda_0 = 0$ and $\gamma > 0$.
For $k = 0, 1, \ldots$ repeat until a stopping criterion is reached:

$x^{(k+1)} = (H^{\mathrm{T}}H + L^{\mathrm{T}}L + I)^{-1}\big(H^{\mathrm{T}}(y_1^{(k)} - q_1^{(k)}) + L^{\mathrm{T}}(y_2^{(k)} - q_2^{(k)}) + (y_3^{(k)} - q_3^{(k)})\big)$,
$a^{(k+1)} = q_1^{(k)} + Hx^{(k+1)}$,
If $a^{(k+1)} > 0$ and $D(b, a^{(k+1)}) \leq \tau$, then
  $\lambda_{k+1} = 0$,
  $y_1^{(k+1)} = a^{(k+1)}$,
Otherwise
  Find λ_{k+1} as solution of $D\left(b, g\left(a^{(k+1)}, \frac{\lambda}{\gamma}\right)\right) = \tau$ by Newton's method initialized by λ_k,
  $y_1^{(k+1)} = g\left(a^{(k+1)}, \frac{\lambda_{k+1}}{\gamma}\right)$,
$y_2^{(k+1)} = \big(I - P_{B_{\|\cdot\|_*}(1/\gamma)}\big)\big(q_2^{(k)} + Lx^{(k+1)}\big)$,
$y_3^{(k+1)} = P_{\mathbb{R}_{\geq 0}}\big(q_3^{(k)} + x^{(k+1)}\big)$,
$q_1^{(k+1)} = a^{(k+1)} - y_1^{(k+1)}$,
$q_2^{(k+1)} = q_2^{(k)} + Lx^{(k+1)} - y_2^{(k+1)}$,
$q_3^{(k+1)} = q_3^{(k)} + x^{(k+1)} - y_3^{(k+1)}$.

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 5.2. *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 II converges to $(\hat{x}, \hat{y}, \hat{q}, \hat{\lambda})$, where $\hat{x}$ is a solution of $(P_{1,\tau})$ and $(P_{2,\hat{\lambda}})$, $\hat{\lambda} > 0$ 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., [26, 29, 52].
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

$$\begin{aligned} g\left(a^{(k)}, \frac{\lambda_k}{\gamma}\right) \;&=\; \frac{1}{2}\Big(a^{(k)} - \frac{1}{\gamma}\,\lambda_k + \sqrt{\big(a^{(k)} - \frac{1}{\gamma}\,\lambda_k\big)^2 + 4\frac{1}{\gamma}b\,\lambda_k}\Big) \;=\; y_1^{(k)}, \\ \Leftrightarrow\; \sqrt{\big(a^{(k)} - \frac{1}{\gamma}\,\lambda_k\big)^2 + 4\frac{1}{\gamma}b\,\lambda_k} \;&=\; 2y_1^{(k)} - \big(a^{(k)} - \frac{1}{\gamma}\,\lambda_k\big), \\ \Leftrightarrow\; \frac{1}{\gamma}\,\lambda_k(b - y_1^{(k)}) \;&=\; y_1^{(k)}(y_1^{(k)} - a^{(k)}), \\ \Leftrightarrow\; \lambda_k(b - y_1^{(k)}) \;&=\; -y_1^{(k)}p_1^{(k)}, \quad p_1^{(k)} := \gamma q_1^{(k)}. \end{aligned} \tag{26}$$

Note that $g(a^{(k)}, 0) = a^{(k)}$, $a^{(k)} > 0$ is also contained in this setting. By Theorem 4.2iii) 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 $i \in \{1, \ldots, n\}$. Thus, we see in the ith equation in (26) 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, (22) implies that $\hat{p}_2 = \gamma\hat{q}_2 \in \partial\|L\hat{x}\|$ and $\hat{p}_3 = \gamma\hat{q}_3 \in N_{\mathbb{R}^n_{\geq 0}}(\hat{x})$ with $H^*\hat{p}_1 + L^*\hat{p}_2 + \hat{p}_3 = 0$. Moreover, we have by (26) and since $H\hat{x} > 0$ that

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

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 (7) 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 (23). Therefore, they are also solutions of $(P_{2,\hat{\lambda}})$ and $(D_{2,\hat{\lambda}})$, respectively. □

5.3 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 (25) 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 [65] 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:

$$\begin{aligned} x^{(k+1)} &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ f_1(x) + \langle p^{(k)}, Ax\rangle + \frac{1}{2s}\|x - x^{(k)}\|_2^2 \right\} \\ &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ \frac{1}{2}\|x - (x^{(k)} - sA^*p^{(k)})\|_2^2 + sf_1(x) \right\}, \\ p^{(k+1)} &= \operatorname*{argmin}_{p\in\mathbb{R}^d} \left\{ f_2^*(p) - \langle p, Ax^{(k+1)}\rangle + \frac{1}{2t}\|p - p^{(k)}\|_2^2 \right\} \\ &= \operatorname*{argmin}_{p\in\mathbb{R}^d} \left\{ \frac{1}{2}\|p - (p^{(k)} + tAx^{(k+1)})\|_2^2 + tf_2^*(p) \right\}. \end{aligned}$$

For our setting (21) 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 [16, 65]:

$$y^{(k+1)} = \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ f_2(y) + \frac{t}{2}\|\frac{1}{t}p^{(k)} + Ax^{(k+1)} - y\|_2^2 \right\}, \tag{27}$$

$$p^{(k+1)} = p^{(k)} + t(Ax^{(k+1)} - y^{(k+1)}). \tag{28}$$

For f_2 as in our setting (21) 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 [47]. The convergence of the algorithm was proved in [16] (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 III (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:

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

Another algorithm which resembles in some way the dual method in [30] was proposed for solving problem (1)/(2) in [61]. The method in [30] uses a predictor-corrector scheme [18] in the alternating direction iterations for the dual variable. This algorithm can be adapted to our setting (21) 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:

$$p^{(k+\frac{1}{2})} = \underset{p\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}\|p - (p^{(k)} + tAx^{(k)})\|_2^2 + tf_2^*(p) \right\}$$
$$x^{(k+1)} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|x - (x^{(k)} - sA^*p^{(k+\frac{1}{2})})\|_2^2 + sf_1(x) \right\},$$
$$p^{(k+1)} = \underset{p\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}\|p - (p^{(k)} + tAx^{(k+1)})\|_2^2 + tf_2^*(p) \right\}.$$

Note that the update steps for p can be splitted again as in (27)-(28).

This algorithm is efficient in the special case when $H = I$ is the identity matrix, e.g., for denoising problems in imaging. Instead of (21) the simpler constraint problem

$$\underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \{\|Lx\| \quad \text{subject to} \quad D(b, x) \le \tau\}, \quad \tau > 0 \tag{29}$$

has to be solved, which can be rewritten as

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

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

Algorithm IV (ADM with predictor-corrector step for minimizing (29))
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:

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

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 Algorithm I as follows:

$h^{(k+1)} = x^{(k)} - s\, L^{\mathrm{T}}p^{(k+\frac{1}{2})}$,
If $h^{(k+1)} > 0$ and $D(b, h^{(k+1)}) \le \tau$, 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 [61].

6 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., [3, 40, 42, 55]. 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. To see this, 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 6.1. 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 $B_t := t\,V$. 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 [64]. 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

$$\begin{aligned}
\mathbb{E}(V \log V) &= \frac{K^K}{\Gamma(K)} \int_0^\infty v^k \,\log v\, \exp(-Kv)\, dv \\
&= \psi(K+1) - \log(K)
\end{aligned}$$

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

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

□

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

Theorem 6.2. *Let $B = (B_1, \ldots, B_n)$ be a random vector and $t = (t_1, \ldots, t_n) \in \mathbb{R}^n_{>0}$.*

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 6.3. *Using standard results for the digamma function ψ, see, e.g., [33, 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 $K \to \infty$.*

7 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 7.1, or the discrete gradient operator

$$\nabla := \begin{pmatrix} I \otimes D \\ D \otimes I \end{pmatrix}, \quad D := \begin{pmatrix} -1 & 1 & & & \\ 0 & -1 & 1 & & \\ & \ddots & \ddots & & \\ & & & -1 & 1 \\ & & & & 0 \end{pmatrix} \tag{30}$$

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 ratios (PSNRs) 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.

7.1 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 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 6. 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 II. As a by-product of the algorithm we obtain by Theorem 5.2 that the penalized problem (17) yields the same solution for the 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 II for the penalized problem with fixed λ we see on the right that for our constrained problem Algorithm II is only slightly slower for equal values of γ.

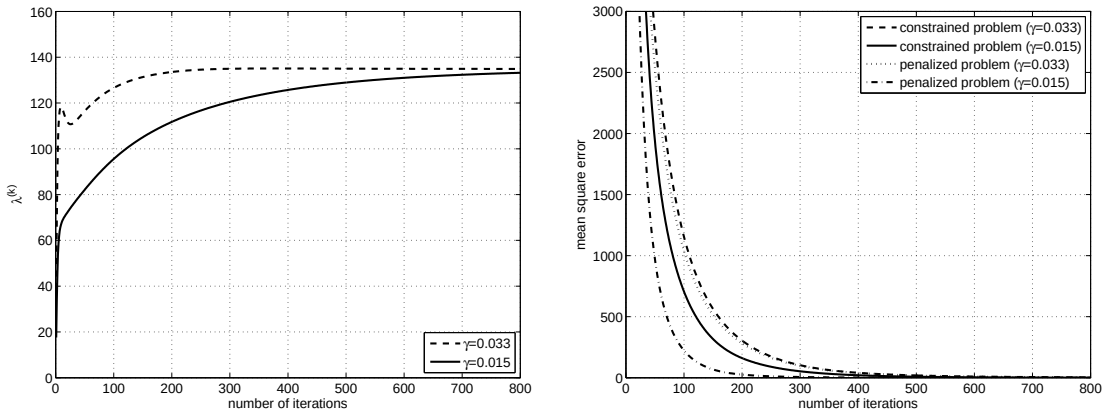


Figure 2: Convergence speed of $x^{(k)}$ and $\lambda^{(k)}$ in Algorithm II when computing the restored image in Figure 1 (right). *Left:* Iterates $\lambda^{(k)}$ for different parameters γ. *Right:* Evolution of the mean square errors $\frac{1}{N}\|x^{(k)} - x^*\|_2^2$ between the intermediate results $x^{(k)}$ and a sufficiently converged reference result x^*.

7.2 Denoising Facing Multiplicative Gamma Noise

TV Regularization Our second example in Figure 3 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 $\tau = \big(\sum_{i=1}^n b_i\big)\big(\psi(K+1) - \log(K)\big) \approx 2.64\,n$ is depicted on the right. For computing this solution we used again Algorithm II with a CG method for solving the occurring linear system of equations. Since H is the identity matrix here, 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

PSNR: 22.63 PSNR: 27.03

Figure 3: *Left:* Original image of the French city of Nîmes (512×512) with values in the range [1, 256], see also [25]. *Middle:* Image corrupted by multiplicative Gamma noise ($K = 10$). *Right:* Restoration result by the I-divergence constrained model (16) with total variation seminorm ($\gamma = 0.015$).

III, where $s\,t < 1/\|L\|_2^2$ is guaranteed for $s\,t < 1/8$. Alternatively, we can also use the predictor-corrector ADM Algorithm IV, here. In Table 1 a speed comparison between these algorithms for 'trial and error' optimized parameters γ, s and t with respect to $\|x^{(k)} - x^*\|_\infty$ is provided, where x^* is a reference result obtained after sufficiently converged Algorithm IV. As the comparison shows Algorithm III is fastest here followed by Algorithm IV if we optimize s and t *disregarding the theoretical convergence constraints* $s\,t < 1/\|L\|_2^2$ and $s\,t < 0.5/\|L\|_2^2$, respectively. For the non-optimized values $s = 1/16$ and $t = 1$ used in [61] Algorithm IV performs worse.
The ADMM Algorithm II is slightly slower than Algorithms III and IV 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 [11, 36] 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 II 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 [31, 55, 56]:

Remark 7.1. *We start with a zero weight matrix $w \in \mathbb{R}^{N,N}$. For every image pixel i we*

Algorithms	Parameters γ	s	t	Number of iterations	Computation times
				$\|x^{(k)} - x^*\|_\infty < 3$	
Algorithm II: ADMM	0.042	–	–	36	2.5 sec
Algorithm II: ADMM with fixed $\lambda = 3.2286$	0.035	–	–	33	1.6 sec
Algorithm III: PDHGMp	–	6.4	0.06	32	1.0 sec
Algorithm IV: ADM with predictor-corr. step	–	3.18	$\frac{1}{17}$	70	2.1 sec
"	–	(1)	$\left(\frac{1}{16}\right)$	(222)	(6.2 sec)
				$\|x^{(k)} - x^*\|_\infty < 1$	
Algorithm II: ADMM	0.055	–	–	66	4.9 sec
Algorithm II: ADMM with fixed $\lambda = 3.2286$	0.058	–	–	67	3.9 sec
Algorithm III: PDHGMp	–	5.4	0.08	52	1.4 sec
Algorithm IV: ADM with predictor-corr. step	–	3.06	$\frac{1}{17}$	95	2.8 sec
"	–	(1)	$\left(\frac{1}{16}\right)$	(284)	(7.9 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 3. The times are averaged here over 100 runs of the algorithms.

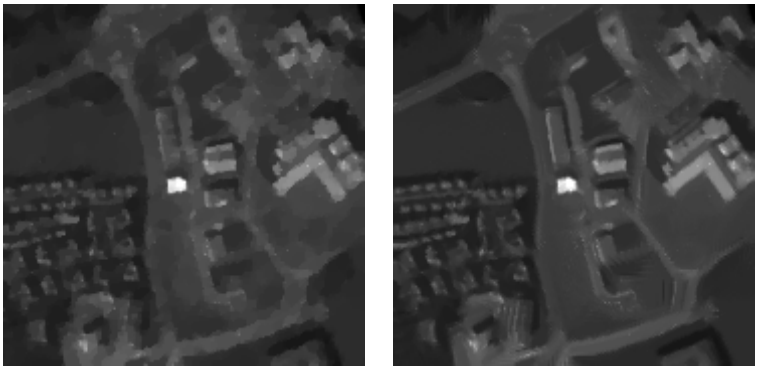
compute for all j within a search window of size $\omega \times \omega$ around i the distances

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

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 $k \le \widetilde{m}$ '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 $L \in \mathbb{R}^{mN,N}$ with $m = 2\widetilde{m}$ so that L consists of m 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 these matrices the constrained problem (16) with the estimated bound $\tau = 2.64\,n$ leads to even better restoration results than the total variation seminorm, see Figure 4.
Table 2 shows a time comparison of the algorithms for solving problem (16) with these nonlocal matrices. The PDHGMp Algorithm III 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, here.

PSNR: 21.79



PSNR: 26.53 PSNR: 26.77

Figure 4: *Top:* Parts of the images depicted in Fig. 3 (left and middle). *Bottom:* Restored images by the I-divergence constrained model (16) with total variation seminorm (left) and a nonlocal term (right), respectively.

8 Conclusions

In this paper we have proposed primal-dual algorithms for solving I-divergence constrained minimization problems. The main 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. However, our minimization algorithms implicitly involve I-divergence penalized problems in the following sense:

- Our algorithm requires the solution of an I-divergence constrained least squares problem in each iteration step. To solve this problem we remember that the solution $\hat{t}(\lambda)$ of a least squares problem with penalized I-divergence is given analytically and depends of course on the regularization parameter λ. Fortunately, there exists a unique parameter λ such that $f(\lambda) := D(b, \hat{t}(\lambda)) = \tau$. Given τ, we first compute this λ by Newton's method and compute $\hat{t}(\lambda)$ afterwards using its analytical expression.

- Despite a solution $\hat{x}$ of the I-divergence constrained problem our algorithm produces a regularization parameter for the penalized problem such that this problem has the same solution, i.e., fulfills a discrepancy principle $D(b, H\hat{x}) = \tau$.

Algorithms	Parameters γ	s	t	Number of iterations	Computation times
				$\|x^{(k)} - x^*\|_\infty < 3$	
Algorithm II: ADMM	0.044	–	–	34	1.27 sec
Algorithm II: ADMM with fixed $\lambda = 6.4454$	0.029	–	–	23	0.75 sec
Algorithm III: PDHGMp	–	2.1	$\frac{1}{17}$	58	0.94 sec
Algorithm IV: ADM with predictor-corr. step	–	4	$\frac{1}{80}$	30	1.05 sec
				$\|x^{(k)} - x^*\|_\infty < 1$	
Algorithm II: ADMM	0.047	–	–	46	1.50 sec
Algorithm II: ADMM with fixed $\lambda = 6.4454$	0.035	–	–	44	1.09 sec
Algorithm III: PDHGMp	–	1.5	$\frac{1}{12}$	95	1.46 sec
Algorithm IV: ADM with predictor-corr. step	–	2.4	$\frac{1}{47}$	67	2.28 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 4 (right). The times are averaged here over 100 runs of the algorithms.

Future directions of research may include the modification of our approach to spatially adapted regularization parameter selection, see [17, 23, 37], and the application of multiplicative iterative update rules for incorporating the non-negativity constraint, cf. [5, 22]. For the first task, further estimates of appropriate parameters τ will be useful. Moreover, the determination of the parameters inherent in the algorithms, i.e., γ and s, t is ongoing research.

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,

$$\lim_{j\to\infty} D(b, Hx_1^{(n_j)}) = D(b, H\hat{x}_1) = \tau_0. \tag{31}$$

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 [48, 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 (31) 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 4. The first lemma is a generalization of a lemma from [19].

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 $\operatorname{lev}_\alpha f$, $\operatorname{lev}_\beta g$ 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 $\operatorname{dom} F \cap \operatorname{dom} G \cap \operatorname{dom} J \neq \emptyset$, then $F + G + J$ attains its finite minimum.

Proof. Since f, g are proper, convex and lower semi-continuous and $\operatorname{lev}_\alpha(f)$, $\operatorname{lev}_\beta(g)$ 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. [48, 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 $\operatorname{dom} F \cap \operatorname{dom} G \cap \operatorname{dom} J \neq \emptyset$, there exist $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma} \in \mathbb{R}$ with $\operatorname{lev}_{\tilde{\alpha}}(F) \cap \operatorname{lev}_{\tilde{\beta}}(G) \cap \operatorname{lev}_{\tilde{\gamma}}(J) \neq \emptyset$. Following the same arguments as in [19, Lemma 3.1 i)] we obtain by $U_2 \cap V_2 = \{0\}$ and the boundedness of $\operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(f)$ and $\operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(g)$ that $\operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(F) \cap \operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(G)$ is bounded. Since $F, G \geq 0$, the level set $\operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(F + G) \subseteq \operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(F) \cap \operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(G)$ is bounded as well and non-empty due to the fact that $\operatorname{lev}_{\tilde{\alpha}+\tilde{\beta}}(F+G) \supseteq \operatorname{lev}_{\tilde{\alpha}}(F) \cap \operatorname{lev}_{\tilde{\beta}}(G) \neq \emptyset$. Since $F + G$ is proper, convex and lower semi-continuous, this implies by [48, 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 $\operatorname{dom} F \cap \operatorname{dom} G \cap \operatorname{dom} J \neq \emptyset$ 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 [49, Thm. 1.9] that $F + G + J$ attains its finite minimum. □

The next lemma is taken from [19].

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 $\operatorname{dom} F \cap \operatorname{dom} G \neq \emptyset$. Then all $\hat{x}, \tilde{x} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \{F(x) + G(x)\}$ 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

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

and applying the definition of the conjugate function this becomes

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

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

$$\begin{aligned} \hat{p} &\in \operatorname*{argmin}_{p \in \mathbb{R}^d} \left\{ f_2^*(p) - \min_{x \in \mathbb{R}^n} \{f_1(x) + \langle p, Ax \rangle\} \right\} \\ &= \operatorname*{argmin}_{p \in \mathbb{R}^d} \left\{ f_2^*(p) + \max_{x \in \mathbb{R}^n} \{\langle -A^* p, x \rangle - f_1(x)\} \right\} \\ &= \operatorname*{argmin}_{p \in \mathbb{R}^d} \{f_2^*(p) + f_1^*(-A^* p)\}. \end{aligned} \tag{32}$$

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

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

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

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

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

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

[9] 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.

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

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

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

[32] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkinsw University Press, Baltimore, London, 1996.

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

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

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

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

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

[42] 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.

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

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

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

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

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

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

[53] 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.

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

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

[56] 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.

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.