Electronic Transactions on Numerical Analysis.
Volume 63, pp. 33–62, 2025.

ISSN 1068–9613.
DOI: 10.1553/etna_vol63s33

ETNA
Kent State University and
Johann Radon Institute (RICAM)

VARIATIONAL POISSON DENOISING VIA AUGMENTED LAGRANGIAN METHODS*

CHRISTIAN KANZOW†, FABIUS KRÄMER‡, PATRICK MEHLITZ§, GERD WACHSMUTH¶, AND FRANK WERNER‖

Abstract. In this paper, we denoise a given noisy image by minimizing a smoothness-promoting function over a set of local similarity measures which compare the mean of the given image and some candidate image on a large collection of subboxes. The associated convex optimization problem possesses a huge number of constraints which are induced by extended real-valued functions stemming from the Kullback–Leibler divergence. Alternatively, these nonlinear constraints can be reformulated as affine ones, which makes the model seemingly more tractable. For the numerical treatment of both formulations of the model (i.e., the original one as well as the one with affine constraints), we propose a rather general augmented Lagrangian method which is capable of handling the huge amount of constraints. A self-contained, derivative-free, global convergence theory is provided, allowing an extension to other problem classes. For the solution of the resulting subproblems in the setting of our suggested image denoising models, we make use of a suitable stochastic gradient method. Results of several numerical experiments are presented in order to compare both formulations and the associated augmented Lagrangian methods.

Key words. augmented Lagrangian method, nonsmooth optimization, Poisson denoising

AMS subject classifications. 49M37, 90C30, 90C48, 90C90

1. Introduction.

Denoising of images is an important task in many applications, which has received considerable attention during the last decades; see, e.g., [25, 49] for recent reviews. In this paper, we consider the problem of estimating an image $\hat{u} \in L^2(\Omega)$, $\hat{u} \geq 0$, on the unit square $\Omega := (0,1)^2$ from a random set of discrete observations $\{\omega_1, \ldots, \omega_N\} \subset \Omega$ with $N \in \mathbb{N}$ related to $\hat{u}$ as follows: We denote by

$$Z := \sum_{i=1}^{N} \delta_{\omega_i},$$

with δ_ω being the Dirac measure centered at $\omega \in \Omega$, the corresponding empirical process and assume that Z is in fact a Poisson point process with intensity $\hat{u}$, i.e., $N \in \mathbb{N}$ is random and

1. for each measurable set $A \subset \Omega$, it holds $\mathbb{E}[Z(A)] = \int_A \hat{u}(\omega)\,\mathrm{d}\omega$, and
2. whenever $A_1, \ldots, A_\ell \subset \Omega$ are measurable and pairwise disjoint, then the random variables $Z(A_1), \ldots, Z(A_\ell)$ are stochastically independent.

*Received June 21, 2024. Accepted December 6, 2024. Published online on January 28, 2025. Recommended by Elena Resmerita. The authors sincerely thank an anonymous reviewer of an earlier version of this paper who suggested the investigation of the reformulation ($\mathrm{VPD}_{\mathrm{aff}}$). The research of Christian Kanzow and Gerd Wachsmuth was supported by the German Research Foundation (DFG) within the priority program "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization" (SPP 1962) under grant numbers KA 1296/24-2 and WA 3636/4-2, respectively.

†University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany (christian.kanzow@uni-wuerzburg.de). ORCID: 0000-0003-2897-2509.

‡University of Regensburg, Faculty of Mathematics, 93040 Regensburg, Germany (Fabius.Kraemer@mathematik.uni-regensburg.de). ORCID: 0009-0005-1422-8502.

§Corresponding Author. Philipps-Universität Marburg, Department of Mathematics and Computer Science, 35032 Marburg, Germany (mehlitz@uni-marburg.de). ORCID: 0000-0002-9355-850X.

¶Brandenburgische Technische Universität Cottbus-Senftenberg, Institute of Mathematics, 03046 Cottbus, Germany (wachsmuth@b-tu.de). ORCID: 0000-0002-3098-1503.

‖University of Würzburg, Institute of Mathematics, 97074 Würzburg, Germany (frank.werner@uni-wuerzburg.de). ORCID: 0000-0001-8446-3587.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We refer to [33] for details on Poisson point processes and emphasize that 1 and 2 already imply that for each measurable set $A \subset \Omega$,

$$Z(A) \sim \operatorname{Poi}\left(\int_A \hat{u}(\omega)\,\mathrm{d}\omega\right), \tag{1.1}$$

i.e., the observations are essentially Poisson distributed. By interpreting the L^2-function $\hat{u}$ as the density of a measure $\hat{U}$ with respect to the Lebesgue measure $\mathrm{d}\omega$, i.e., $\hat{U}(A) = \int_A \hat{u}(\omega)\,\mathrm{d}\omega$ for all measurable sets $A \subset \Omega$, the relation (1.1) shows that the measure Z is in fact a noisy version of $\hat{U}$ with $\mathbb{E}[Z(A)] = \hat{U}(A)$ for all measurable $A \subset \Omega$ due to 1. In view of the Radon–Nikodým theorem, there is a one-to-one correspondence between $\hat{u}$ and the absolutely continuous measure $\hat{U}$, and hence, Z can be interpreted as a noisy (in fact Poissonian) version of $\hat{u}$.

As the Poisson distribution is a natural model in applications ranging from astronomy to biophysics (see, e.g., [2, 5, 33, 51]), the problem of estimating $\hat{u}$ from Z has received considerable attention over the past decades. Early references include [47, 48], where explicit models for the noise occurring in Charge Coupled Device cameras and corresponding methods for its removal where discussed. Since then, a major focus has been on variational approaches; see, e.g., [3, 4, 8, 9, 10, 26, 40, 41, 54]. In all these works, a (convex) functional $J\colon L^2(\Omega) \to \overline{\mathbb{R}}$ of the composite form

$$\forall u \in L^2(\Omega)\colon \quad J(u) := G_Z(u) + f(u) \tag{1.2}$$

is minimized, where $G_Z\colon L^2(\Omega) \to \overline{\mathbb{R}}$ is a data-fidelity term measuring the discrepancy between the observations Z and the candidate image $u \in L^2(\Omega)$ and $f\colon L^2(\Omega) \to \overline{\mathbb{R}}$ is a regularization term promoting desired properties of u (e.g., smoothness, sparsity,...). A natural choice for G_Z is the negative log-likelihood functional of the Poisson distribution

$$G_Z(u) := \int_\Omega u(\omega)\,\mathrm{d}\omega - \int_\Omega \ln(u(\omega))\,\mathrm{d}Z(\omega) = \int_\Omega u(\omega)\,\mathrm{d}\omega - \sum_{i=1}^{N} \ln\left(u(\omega_i)\right). \tag{1.3}$$

The previously mentioned works do all rely on this data-fidelity term and mostly differ in the choices of f and the algorithm used for minimization. The smoothness-promoting function f can be chosen depending on the application. Famous choices include classical L^2-norm penalties, sparsity-promoting penalties such as $f(u) := \sum_{i=1}^{\infty} |(u, e^i)_{L^2(\Omega)}|$ with a complete orthonormal system or frame $\{e^i\}_{i\in\mathbb{N}} \subset L^2(\Omega)$, or the TV-seminorm given by

$$\mathrm{TV}(u) := \sup\left\{ \int_\Omega u(\omega)\ \operatorname{div}\varphi(\omega)\,\mathrm{d}\omega \,\middle|\, \varphi \in C_c^1(\Omega;\mathbb{R}^2), \max_{\omega\in\Omega} \|\varphi(\omega)\|_2 \le 1 \right\}$$

for each $u \in L^2(\Omega)$, which equals $\mathrm{TV}(u) = \int_\Omega \|\nabla u(\omega)\|_2\,\mathrm{d}\omega$ for differentiable u; see [1] for details. Above, $C_c^1(\Omega;\mathbb{R}^2)$ represents the space of all continuously differentiable functions mapping from Ω to $\mathbb{R}^2$ with compact support in Ω. In the case where smooth functions are desired, also Sobolev-type penalties

$$f(u) := \int_{\mathbb{R}^2} \left(1 + \|\zeta\|_2^2\right)^\theta |(\mathfrak{F}u)(\zeta)|^2 \mathrm{d}\zeta, \tag{1.4}$$

are suitable, where $\theta \ge 0$ and $\mathfrak{F}u$ denotes the Fourier transform of u extended by 0 to all of $\mathbb{R}^2$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

1.1. Problem statement. In this paper, we will follow a different approach proposed in [27, 28]. It is based on considering a constrained problem

$$\min_{u\in L^2(\Omega)} f(u) \qquad \text{s.t.} \quad H_Z(u) \le c \tag{1.5}$$

for some function $H_Z\colon L^2(\Omega) \to \overline{\mathbb{R}}^m$ depending on Z and some vector $c \in \mathbb{R}^m$ instead of the problem $\min_{u\in L^2(\Omega)} J(u)$ with J as in (1.2). Note that for $m = 1$ and $H_Z = \sigma G_Z$, with some $\sigma > 0$, this problem is in fact related to $\min_{u\in L^2(\Omega)} J(u)$; see [55]. However, motivated by results from multiscale statistics, we employ a different choice than (simply) the negative log-likelihood term from (1.3) for the constraint function H_Z in (1.5) and aim to minimize f only over images which are everywhere locally compatible to Z. For the latter, we introduce a (carefully chosen) finite system $\mathcal{B} \subset 2^\Omega$ of measurable regions with positive measure in Ω (e.g., a set of square subboxes of the image) and consider a candidate image $u \in L^2(\Omega)$ as compatible with the data if and only if its mean $u_B := |B|^{-1}\int_B u(\omega)\,\mathrm{d}\omega$ with the Lebesgue measure $|B|$ of B does not deviate too much from the mean $Z_B := |B|^{-1}Z(B)$ of the data Z on B for all $B \in \mathcal{B}$. Given the Poisson distribution of Z_B, the deviation of u_B from Z_B can be made precise by means of statistical hypothesis testing or, as a specific instance, by the local likelihood ratio test (LRT for short) statistic

$$\forall u \in L^2(\Omega)\colon \quad T_B(Z,u) := \sqrt{2\,|B|\,(u_B - Z_B + Z_B \ln(Z_B/u_B))}. \tag{1.6}$$

Whenever the local LRT statistic $T_B(Z,u)$ is too large (which can be made precise when specifying the type-1 error of the LRT), the candidate image u is considered incompatible with Z on B.

This motivates the consideration of the *variational Poisson denoising* optimization problem

$$\min_{u\in L^2(\Omega)} f(u) \qquad \text{s.t.} \quad \eta(Z_B, u_B) \le r(|B|) \quad \forall B \in \mathcal{B}, \tag{VPD}$$

with a function $r\colon [0,1] \to (0,\infty)$ reflecting that the right-hand side of the constraints should—similar to the potential number of possible regions—depend on the *scale* $|B|$ only and the so-called *Kullback–Leibler divergence* $\eta\colon \mathbb{R}^2 \to \overline{\mathbb{R}}$ given by

$$\forall (a,b) \in \mathbb{R}^2\colon \quad \eta(a,b) := \begin{cases} b - a + a\ln(a/b) & \text{if } a > 0,\ b > 0,\\ b & \text{if } a = 0,\ b \ge 0,\\ \infty & \text{otherwise.}\end{cases} \tag{1.7}$$

Note that η is a nonnegative, convex, and lower semicontinuous function that is continuously differentiable on $\{(a,b) \in \mathbb{R}^2 \,|\, a,b > 0\}$; see, e.g., [33, 50]. However, η is discontinuous precisely at the points from $\{(a,b) \in \mathbb{R}^2 \,|\, a,b \ge 0,\ ab = 0\}$ and thus essentially nonsmooth. Furthermore, note the similarity to the negative log-likelihood term G_Z in (1.3), which differs from an integral over η only by terms independent of the reconstruction candidate u.

The variational Poisson denoising problem is methodologically appealing from a statistical point of view and has a very important property which is not directly obtained for the variational or any other method mentioned before. If the function r is chosen such that

$$\mathbb{P}\left[\forall B \in \mathcal{B}\colon\ \eta(Z_B, \hat{u}_B) \le r(|B|)\right] \ge \alpha \tag{1.8}$$

holds for the true image $\hat{u}$, i.e., if 0 is a $(1-\alpha)$-quantile of the random variable given by $\sup_{B\in\mathcal{B}}\left[\eta(Z_B,\hat{u}_B) - r(|B|)\right]$ (see [38]), then each reconstruction $\bar{u} \in L^2(\Omega)$ solving (VPD) automatically satisfies

$$\mathbb{P}\left[f(\bar{u}) \le f(\hat{u})\right] \ge \alpha, \tag{1.9}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

i.e., with probability at least α, the reconstruction $\bar{u}$ is at least as smooth as the true image $\hat{u}$. Further theoretical properties of a similar method in the case of Gaussian observations such as optimality for given function classes have been considered in [19, 29]. From a practical perspective, (1.8) gives a clear interpretation on how to choose r, and it ensures that all remaining parameters of the method have a direct interpretation (such as the probability α). However, this property comes at the price that (VPD) is a computationally demanding nonsmooth convex problem, which especially suffers from a huge number of constraints. For example, for the setting we will consider in our numerical examples below where u is discretized by means of 266^2 equally sized pixels, then choosing $\mathcal{B} \subset 2^\Omega$ as the family of all subsquares in the image with side length (scale) between 1 and 64 pixels leads already to 3 541 216 constraints; see Section 3.2.

In [27, 28], the problem (VPD) has been tackled by an ADMM approach. This requires that in each iteration, a projection to the feasible set

$$\begin{aligned}\mathcal{F} &:= \left\{u \in L^2(\Omega) \,\middle|\, \forall B \in \mathcal{B}\colon\ \eta(Z_B, u_B) \le r(|B|)\right\} \\ &= \bigcap_{B \in \mathcal{B}} \left\{u \in L^2(\Omega) \,\middle|\, \eta(Z_B, u_B) \le r(|B|)\right\}\end{aligned}$$

has to be computed. In the previously mentioned works, this has been carried out by Dykstra's algorithm, leading to a computationally poor performance. Let us note that if the noisy image Z possesses L^2-regularity, then it would be feasible to solve (VPD).

Further implementations of comparable problems (in the Gaussian case) have been discussed in [20, 43], relying on different algorithmic frameworks such as the Chambolle–Pock algorithm (see [14]) or semismooth Newton methods (see [32]), but these approaches explicitly exploit the Gaussian and hence the smooth structure of the corresponding optimization problem.

1.2. Augmented Lagrangian methods. In this paper, we will approach (VPD) by means of augmented Lagrangian methods, which provide a well-established framework for the numerical solution of constrained optimization problems; see, e.g., [6, 7]. The principal idea behind those methods is to replace the minimization of a function subject to difficult constraints by the minimization of the sum of the associated Lagrangian function and a quadratic penalty term. Typically, the arising subproblems are unconstrained or at least possess merely simple constraints. In contrast to penalty methods, convergence results for augmented Lagrangian methods do not necessarily require that the penalty parameter tends to infinity. The augmented Lagrangian method should be viewed as a general framework which allows an adaptation to many different scenarios simply by taking a suitable and problem-dependent subproblem solver. The two standard references mentioned above consider the situation of a general nonlinear program (in finite dimensions), but a suitable (global) convergence theory tailored for appropriate stationary points is also available for some difficult, structured, and/or nonsmooth problems. This includes situations with an abstract geometric constraint (with a potentially nonconvex constraint set) (see [30, 35]) and programs with a composite objective function (see [15, 17, 18, 21, 31, 39, 45]). Specifically, in [45], issues of nonsmoothness are eliminated by exploiting smoothness properties of the Moreau envelope in a partially convex situation. The fully nonsmooth setting is also discussed in [24, 52], where all functions are smoothed, as well as in [42] in the framework of so-called difference-of-convex programs.

While these references mainly deal with finite-dimensional problems, the augmented Lagrangian approach can also be extended to the infinite-dimensional situation. Here, we distinguish between the "half" and "full" infinite-dimensional setting. Both settings share the property that the optimization variables belong to a Banach space, but the former allows

ETNA
Kent State University and
Johann Radon Institute (RICAM)

only finitely many inequality constraints (possibly additional equality as well as abstract constraints), whereas the latter allows more general functional constraints stated in a Banach space (say, $G(x) \in K$ for a mapping $G\colon X \to Y$ between two Banach spaces X and Y and a closed, convex set $K \subset Y$). The convergence theory for the "half" infinite-dimensional setting was already considered in the seminal paper [44] by Rockafellar; see also the monograph [34]. Extensions to the fully infinite-dimensional setting are given in [12, 13, 37].

We should note, however, that there exist different versions for a realization of the augmented Lagrangian approach. In particular, there is the classical method with the standard Hestenes–Powell–Rockafellar update of the Lagrange multipliers, and there is the safeguarded version with a more careful updating of the multiplier estimates; see [7]. On the one hand, the counterexample in [36] shows that there cannot exist a satisfactory global convergence theory for the classical method, at least not in the nonconvex setting, while the existing convergence theory for the safeguarded version is rather complete in the sense that it has all desirable (and realistic) properties. On the other hand, for convex problems, there exists a convergence theory for the classical approach even with a constant penalty parameter. This result was established by Rockafellar in [44], even for the "half" infinite-dimensional setting, and is based on the duality of the augmented Lagrangian and the proximal point method; see [45] as well. In particular, this convergence theory is based on the existence of Lagrange multipliers.

1.3. Our contributions. As mentioned earlier, the purpose of this article is to study the numerical solution of (VPD) with the aid of augmented Lagrangian methods. Let us note that (VPD) is covered by the "half" infinite-dimensional setting, and due to the huge number of constraints in (VPD), the augmentation approach seems to be perfectly suitable to tackle the problem computationally. In the course of the paper, we also suggest a reformulation of the constraints in (VPD) as purely affine inequalities, and the latter is covered by the "half" infinite-dimensional setting as well. Though both models are convex, we provide a purely primal convergence theory for a safeguarded augmented Lagrangian method in a more general nonconvex setting. We assume, however, that we are able to find an approximate global minimum of the resulting subproblems. This is a realistic scenario for the convex optimization problem (VPD) and its aforementioned reformulation but might also be applicable in some other situations (e.g., think of disjunctive constraint systems composed of finitely many convex branches). In contrast to the existing literature, our analysis is derivative-free and works under minimal continuity assumptions on the data. This is rather important as the constraints in (VPD) are modeled with the aid of the discontinuous function η from (1.7). Moreover, we stress that our convergence theory is independent of any assumption regarding the existence (or uniqueness or boundedness) of Lagrange multipliers. Finally, let us mention that the method has favorable convergence properties even in the case where the constraints are inconsistent since it still provides limits which minimize a suitable measure for the constraint violation.

The suggested safeguarded augmented Lagrangian method is then applied to the numerical solution of (VPD) as well as to its reformulation with affine constraints. As the latter is guaranteed to possess Lagrange multipliers at its minimizers, we also apply to it the classical augmented Lagrangian method (i.e., we abstain from safeguarding) with a constant penalty parameter which is reasonable due to the analysis in [44]. We compare these three numerical approaches by means of several different test instances and document the results.

The paper is organized in the following way: In the remainder of this introductory section, we comment on the notation that will be used throughout the paper. The safeguarded augmented Lagrangian method of our interest is stated and analyzed in Section 2, where we consider nonsmooth problems with finitely many inequality constraints, a general operator equation (representing, e.g., a partial differential equation), as well as an abstract constraint set such that the associated augmented Lagrangian subproblems can be solved up to approximate

ETNA
Kent State University and
Johann Radon Institute (RICAM)

global optimality. Section 3 is dedicated to the numerical solution of the Poisson denoising model with the aid of augmented Lagrangian methods. In Section 3.1, we derive a reformulation of (VPD) which merely possesses (finitely many) affine inequality constraints. The implementation of our experiments, the exploited subproblem solver, and the way we are documenting our results are discussed in Sections 3.2–3.4 , respectively, before in Section 3.5 the numerical performance of augmented Lagrangian methods when applied to (VPD) and its reformulation is illustrated and compared based on several different test instances. We conclude with some final remarks in Section 4.

1.4. Notation. Let $\mathbb{R}$ and $\mathbb{R}_+$ denote the sets of all real numbers and nonnegative real numbers, respectively. We make use of $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. Throughout the paper, for a given finite set D, $\#D$ is used to denote the cardinality of D. Let $n \in \mathbb{N}$ be a positive integer. For vectors $x, y \in \mathbb{R}^n$, $\max(x, y) \in \mathbb{R}^n$ denotes the componentwise maximum of x and y. For any $p \in [1, \infty]$, the ℓ_p-norm of $x \in \mathbb{R}^n$ will be denoted by $\|x\|_p$.

Whenever X is a Banach space, its norm will be denoted by $\|\cdot\|_X : X \to [0, \infty)$ if not stated otherwise. Strong and weak convergence of a sequence $\{x^k\} \subset X$ to $x \in X$ are represented by $x^k \to x$ and $x^k \rightharpoonup x$, respectively. If $K \subset \mathbb{N}$ is a set of infinite cardinality, then we make use of $x^k \to_K x$ ($x^k \rightharpoonup_K x$) in order to express that the subsequence $\{x^k\}_{k \in K}$ converges (converges weakly) to x as k tends to ∞ in K (which we denote by $k \to_K \infty$ for brevity). The (topological) dual space of X will be represented by X^*, and the associated dual pairing is then denoted by $\langle \cdot, \cdot \rangle_X : X^* \times X \to \mathbb{R}$. Let Y be another Banach space. If $h \colon X \to Y$ is Fréchet differentiable at $x \in X$, $h'(x) \colon X \to Y$ denotes the derivative of h at x. Similarly, if X_1 and X_2 are Banach spaces such that $X = X_1 \times X_2$, and if h is Fréchet differentiable at $x := (x_1, x_2) \in X$, $h'_{x_1}(x) \colon X_1 \to Y$ denotes the partial derivative with respect to x_1 of h at x. The inner product in a Hilbert space H will be represented by $(\cdot, \cdot)_H \colon H \times H \to \mathbb{R}$.

For an arbitrary function $\varphi \colon X \to \overline{\mathbb{R}}$ defined on a Banach space X, the set $\operatorname{dom} \varphi := \{x \in X \,|\, \varphi(x) < \infty\}$ is referred to as the domain of φ. Whenever φ is convex and $\bar{x} \in \operatorname{dom} \varphi$ is chosen arbitrarily, the set

$$\partial \varphi(\bar{x}) := \{\xi \in X^* \,|\, \forall x \in \operatorname{dom} \varphi \colon \ \varphi(x) \geq \varphi(\bar{x}) + \langle \xi, x - \bar{x} \rangle_X\}$$

is called the subdifferential (in the sense of convex analysis) of φ at $\bar{x}$.

For an integer $d \in \mathbb{N}$, a bounded open set $\Omega \subset \mathbb{R}^d$, and $p \in [1, \infty)$, let $L^p(\Omega)$ denote the Lebesgue space of (equivalence classes of) measurable functions $u \colon \Omega \to \mathbb{R}$ such that $\Omega \ni \omega \mapsto |u(\omega)|^p \in \mathbb{R}$ is integrable, equipped with the standard norm, which we denote by $\|\cdot\|_p : L^p(\Omega) \to [0, \infty)$. Note that it will be clear from the context whether $\|\cdot\|_p$ is taken in $\mathbb{R}^n$ or $L^p(\Omega)$.

2. An augmented Lagrangian method for nonsmooth optimization problems. In this section we address the algorithmic treatment of the rather abstract optimization problem

$$\text{(P)} \qquad \min_{x \in X} \ f(x) \qquad \text{s.t.} \quad g(x) \leq 0, \ h(x) = 0, \ x \in C,$$

where $f \colon X \to \overline{\mathbb{R}}$, $g \colon X \to \overline{\mathbb{R}}^m$, and $h \colon X \to Y$ are given functions and $C \subset X$ is a weakly sequentially closed set. Moreover, X is a reflexive Banach space, and Y is a Hilbert space, which we identify with its dual, i.e., $Y \cong Y^*$. We assume that the functions $f, g_1, \ldots, g_m \colon X \to \overline{\mathbb{R}}$ are weakly sequentially lower semicontinuous, while the function h is weakly-strongly sequentially continuous in the sense that

$$\forall \{x^k\} \subset X, \forall x \in X \colon \quad x^k \rightharpoonup x \quad \text{in } X \quad \Longrightarrow \quad h(x^k) \to h(x) \quad \text{in } Y.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

In particular, h is continuous. In contrast to the standard setting of nonlinear programming, we abstain from demanding any differentiability properties of the data functions. Throughout this section, let $\mathcal{F} \subset X$ denote the feasible set of (P). Note that at least continuity of the function h is indispensable in order to guarantee that $\mathcal{F}$ is closed. For later use, we introduce $\operatorname{dom} g := \bigcap_{i=1}^{m} \operatorname{dom} g_i$. Here, $g_1, \ldots, g_m$ are the component functions of g. As a minimal requirement, we need

$$\operatorname{dom} f \cap \operatorname{dom} g \cap C \neq \emptyset, \tag{2.1}$$

which will be a standing assumption in this section. Observe that (2.1) does not necessarily rule out $\operatorname{dom} f \cap \mathcal{F} = \emptyset$. However, $\operatorname{dom} f \cap \mathcal{F} \neq \emptyset$ clearly implies the validity of (2.1).

The assumptions from above already guarantee that $\mathcal{F}$ is weakly sequentially closed. Together with $\operatorname{dom} f \cap \mathcal{F} \neq \emptyset$ and the weak sequential lower semicontinuity of the objective functional f, this can be interpreted as a minimal requirement in constrained optimization in order to ensure that the underlying optimization problem (P) possesses a solution. This would be inherent whenever $\mathcal{F}$ is, additionally, bounded or f is, additionally, coercive as standard arguments show.

2.1. A chain rule for lower semicontinuity. Before we can start with the presentation of the augmented Lagrangian method and its convergence analysis, we need to prepare conditions that guarantee that the composition of a (weakly sequentially) lower semicontinuous function and a continuous function is (weakly sequentially) lower semicontinuous again. Such a criterion is presented in the following lemma:

LEMMA 2.1. *For some Banach space X, let $\varphi\colon X \to \overline{\mathbb{R}}$ be weakly sequentially lower semicontinuous, and let $\psi\colon \mathbb{R} \to \mathbb{R}$ be a continuous and monotonically increasing function. Then $\psi \circ \varphi\colon X \to \overline{\mathbb{R}}$ defined via*

$$\forall x \in X\colon \quad (\psi \circ \varphi)(x) := \begin{cases} \psi(\varphi(x)) & \text{if } \varphi(x) < \infty, \\ \lim_{t\to\infty} \psi(t) & \text{if } \varphi(x) = \infty \end{cases}$$

is weakly sequentially lower semicontinuous.

Proof. Choose $\{x^k\} \subset X$ and $\bar{x} \in X$ with $x^k \rightharpoonup \bar{x}$ arbitrarily, and pick an infinite set $K \subset \mathbb{N}$ such that

$$\alpha := \liminf_{k\to\infty} (\psi \circ \varphi)(x^k) = \lim_{k\to_K\infty} (\psi \circ \varphi)(x^k).$$

First, we argue that α cannot attain the value $-\infty$. Indeed, weak sequential lower semicontinuity of φ yields the estimate $-\infty < \varphi(\bar{x}) \leq \liminf_{k\to\infty} \varphi(x^k)$, so we infer that $\{\varphi(x^k)\}$ is bounded from below. Consequently, $\{(\psi \circ \varphi)(x^k)\}$ is also bounded from below, and thus, $\alpha > -\infty$. In the case $\alpha = \infty$, we automatically have $\alpha \geq (\psi \circ \varphi)(\bar{x})$, and thus, there is nothing to show. Hence, we assume $\alpha \in \mathbb{R}$. By weak sequential lower semicontinuity of φ, we have $\beta := \liminf_{k\to_K\infty} \varphi(x^k) \geq \varphi(\bar{x})$. Pick an infinite set $K' \subset K$ such that $\lim_{k\to_{K'}\infty} \varphi(x^k) = \beta$. In the case where $\beta \in \mathbb{R}$ holds, $\varphi(\bar{x})$ and the tail of the sequence $\{\varphi(x^k)\}_{k\in K'}$ are finite, so we find

$$\alpha = \lim_{k\to_{K'}\infty} (\psi \circ \varphi)(x^k) = \lim_{k\to_{K'}\infty} \psi(\varphi(x^k)) = \psi(\beta) \geq \psi(\varphi(\bar{x})) = (\psi \circ \varphi)(\bar{x})$$

by continuity and monotonicity of ψ. Next, suppose that $\beta = \infty$ holds. In the case where $\{\varphi(x^k)\}_{k\in K'}$ equals ∞ along the tail of the sequence, we find

$$\alpha = \lim_{k\to_{K'}\infty} (\psi \circ \varphi)(x^k) = \lim_{t\to\infty} \psi(t) \geq (\psi \circ \varphi)(\bar{x})$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

by monotonicity of ψ. Otherwise, there is an infinite set $K'' \subset K'$ such that we have $\{\varphi(x^k)\}_{k\in K''} \subset \mathbb{R}$. However, $\beta = \infty$ yields $\lim_{k\to_{K''}\infty} \varphi(x^k) = \infty$. Hence, by definition of the composition, we find

$$\alpha = \lim_{k\to_{K''}\infty} (\psi\circ\varphi)(x^k) = \lim_{k\to_{K''}\infty} \psi(\varphi(x^k)) = \lim_{t\to\infty} \psi(t) \geq (\psi\circ\varphi)(\bar{x})$$

by continuity and monotonicity of ψ. This completes the proof. ∎

We would like to note that, in general, for a (weakly sequentially) lower semicontinuous function $\varphi\colon X \to \overline{\mathbb{R}}$, the mappings $x \mapsto |\varphi(x)|$ and $x \mapsto \varphi^2(x)$ are not (weakly sequentially) lower semicontinuous (for example, choose $X := \mathbb{R}$, and set $\varphi(x) := -1$ for all $x \leq 0$ and $\varphi(x) := 0$ for all $x > 0$). Observe that the absolute value function and the square are not monotonically increasing, i.e., the assumptions of Lemma 2.1 are not satisfied in this situation.

We comment on a typical setting where Lemma 2.1 applies.

EXAMPLE 2.2. For each $\alpha > 0$ and $\beta \in \mathbb{R}$, the function $\psi\colon \mathbb{R} \to \mathbb{R}$ given by $\psi(t) := \max^2(0, \alpha t + \beta)$ for each $t \in \mathbb{R}$ is continuous, monotonically increasing, and satisfies $\lim_{t\to\infty} \psi(t) = \infty$. Thus, for each Banach space X and each weakly sequentially lower semicontinuous function $\varphi\colon X \to \overline{\mathbb{R}}$, the composition $\psi\circ\varphi\colon X \to \overline{\mathbb{R}}$ given by

$$\forall x \in X\colon \quad (\psi\circ\varphi)(x) := \begin{cases} \psi(\varphi(x)) & \text{if } \varphi(x) < \infty, \\ \infty & \text{if } \varphi(x) = \infty \end{cases}$$

is weakly sequentially lower semicontinuous as well by Lemma 2.1.

We also note that this particular function ψ is convex. Thus, keeping the monotonicity of ψ in mind, whenever φ is convex, the composition $\psi\circ\varphi$ is convex as well.

2.2. Statement of the algorithm. Let $L\colon X \times \mathbb{R}^m_+ \times Y \to \overline{\mathbb{R}}$ denote the Lagrangian function associated with (P), which is given by

$$L(x,\lambda,\mu) := f(x) + \lambda^\top g(x) + (\mu, h(x))_Y$$

for $x \in X$, $\lambda \in \mathbb{R}^m_+$, and $\mu \in Y$. For the construction of our solution method, we make use of the corresponding augmented Lagrangian function $L_\rho\colon X \times \mathbb{R}^m_+ \times Y \to \overline{\mathbb{R}}$ associated with (P) and which is given by

$$\begin{aligned} L_\rho(x,\lambda,\mu) := f(x) &+ \frac{1}{2\rho}\sum_{i=1}^m \left(\max^2(0, \lambda_i + \rho g_i(x)) - \lambda_i^2\right) \\ &+ (\mu, h(x))_Y + \frac{\rho}{2}\|h(x)\|_Y^2, \end{aligned} \tag{2.2}$$

for all $x \in X$, $\lambda \in \mathbb{R}^m_+$, and $\mu \in Y$, where $\rho > 0$ is a given penalty parameter. Roughly speaking, the latter results from L by adding a standard quadratic penalty for the constraints $g(x) \leq 0$ and $h(x) = 0$ appearing in (P). Within our algorithmic framework, the function L_ρ has to be minimized with respect to x, which means that the term $-\frac{1}{2\rho}\|\lambda\|_2^2$ could be removed from the definition of L_ρ. However, for some of the proofs that we are going to provide, it will be beneficial to keep this shift. We would like to point the reader's attention to the fact that the function $L_\rho(\cdot,\lambda,\mu)$ is weakly sequentially lower semicontinuous for each $\lambda \in \mathbb{R}^m_+$ and $\mu \in Y$ due to Lemma 2.1, Example 2.2, and the fact that the function h is weakly-strongly sequentially continuous. Furthermore, we would like to point out that the abstract constraint set C is not incorporated into the definitions of L and L_ρ on purpose.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

REMARK 2.3. Whenever (P) is a convex optimization problem, i.e., whenever the functions $f, g_1, \dots, g_m$ are convex while h is affine, then, for each $\lambda \in \mathbb{R}^m_+$ and $\mu \in Y$, $L_\rho(\cdot, \lambda, \mu)$ is a convex function as well by monotonicity and convexity of $t \mapsto \max^2(0, \alpha t + \beta)$ for each $\alpha > 0$ and $\beta \in \mathbb{R}$.

For some penalty parameter $\rho > 0$, we introduce a function $V_\rho \colon X \times \mathbb{R}^m_+ \to \overline{\mathbb{R}}$ by means of

$$V_\rho(x, \lambda) := \begin{cases} \max\left(\|\max(g(x), -\lambda/\rho)\|_\infty, \|h(x)\|_Y\right) & \text{if } x \in \operatorname{dom} g, \\ \infty & \text{if } x \notin \operatorname{dom} g, \end{cases}$$

for all $x \in X$ and $\lambda \in \mathbb{R}^m_+$. From the definition of V_ρ, we obtain

$$V_\rho(x, \lambda) = 0 \quad \Longleftrightarrow \quad g(x) \leq 0,\ \lambda \geq 0,\ \lambda^\top g(x) = 0,\ h(x) = 0,$$

i.e., V_ρ can be used to measure the feasibility of x for (P) with respect to the constraints induced by g and h as well as the validity of the complementarity-slackness condition with respect to the inequality constraints.

In Algorithm 1, we state a pseudo-code which describes our method.

Algorithm 1 Safeguarded Augmented Lagrangian Method for (P).

Require: nonempty, closed, convex, and bounded sets $B_m \subset \mathbb{R}^m_+$ and $B_Y \subset Y$, starting point $(x^0, \lambda^0, \mu^0) \in C \times \mathbb{R}^m_+ \times Y$, initial penalty parameter $\rho_0 > 0$, parameters $\tau \in (0, 1)$, $\gamma > 1$

1: Set $k := 0$.
2: **while** a suitable termination criterion is violated at iteration k **do**
3: Choose v^k and w^k as the projections of λ^k and μ^k onto B_m and B_Y, respectively.
4: Compute $x^{k+1} \in C$ as an approximate solution of the optimization problem

$$\min_{x \in X} L_{\rho_k}(x, v^k, w^k) \qquad \text{s.t.} \quad x \in C. \tag{2.3}$$

5: Set

$$\lambda^{k+1} := \max\left(0, v^k + \rho_k g(x^{k+1})\right), \qquad \mu^{k+1} := w^k + \rho_k h(x^{k+1}). \tag{2.4}$$

6: If $k = 0$, or if the condition

$$V_{\rho_k}(x^{k+1}, v^k) \leq \tau\, V_{\rho_{k-1}}(x^k, v^{k-1}) \tag{2.5}$$

holds, then set $\rho_{k+1} := \rho_k$; otherwise set $\rho_{k+1} := \gamma \rho_k$.
7: Set $k \leftarrow k + 1$.
8: **end while**
9: **return** x^k

In Algorithm 1, the quantities v^k and w^k play the role of Lagrange multiplier estimates. By construction, the sequences $\{v^k\}$ and $\{w^k\}$ remain bounded throughout a run of the algorithm while this does not necessarily hold true for $\{\lambda^k\}$ and $\{\mu^k\}$. Note that the classical augmented Lagrangian method could be recovered from Algorithm 1 by replacing v^k and w^k by λ^k and μ^k everywhere, respectively, and removing Step 3. However, the so-called

ETNA
Kent State University and
Johann Radon Institute (RICAM)

safeguarded variant from Algorithm 1 has been shown to possess better global convergence properties than the classical method; see, e.g., [36] for details. For example, one can choose B_m as the (very large) box $[0, \mathbf{v}]$ for some $\mathbf{v} \in \mathbb{R}^m$ satisfying $\mathbf{v} > 0$. A similar choice is possible for B_Y if Y is equipped with a suitable partial order relation. This way, Algorithm 1 is likely to parallel the classical augmented Lagrangian method if the sequences $\{\lambda^k\}$ and $\{\mu^k\}$ remain bounded. Let us mention that, in principle, the convergence analysis associated with Algorithm 1, which is presented in Section 2.3, just requires the inclusions $v^k \in B_m$ and $w^k \in B_Y$ in Step 3.

Assuming for a moment that all involved data functions are smooth, then the derivative with respect to x of L_ρ from (2.2) is given by

$$(L_\rho)'_x(x, \lambda, \mu) = f'(x) + \sum_{i=1}^{m} \max(0, \lambda_i + \rho\, g_i(x))\, g_i'(x) + h'(x)^*[\mu + \rho h(x)].$$

Thus, the updating rule for the multipliers in (2.4) yields

$$(L_{\rho_k})'_x(x^{k+1}, v^k, w^k) = L'_x(x^{k+1}, \lambda^{k+1}, \mu^{k+1}), \tag{2.6}$$

which is the basic idea behind Step 5. Note that a similar formula as (2.6) can be obtained in terms of several well-known concepts of subdifferentiation whenever a suitable chain rule applies.

Finally, let us mention that in Step 6, the penalty parameter is increased whenever the new iterate (x^{k+1}, v^k, w^k) is not (sufficiently) better from the viewpoint of feasibility (and complementarity) than the old iterate (x^k, v^{k-1}, w^{k-1}). Note that our choice for the infinity norm in the definition of V_ρ is a matter of taste since all norms are equivalent in finite-dimensional spaces. However, this particular error measure V_ρ keeps track of the largest violation of the feasibility and complementarity condition with respect to *all* inequality constraints, which is why we favor it here.

For further information about (safeguarded) augmented Lagrangian methods in nonlinear programming, we refer the interested reader to the monograph [7].

2.3. Convergence to global minimizers. In this section, we provide a convergence analysis for Algorithm 1, where we assume that in Step 4, the subproblem (2.3) is solved up to (approximate) global optimality. For example, this is possible whenever (P) is a convex program (see Remark 2.3) but also in more general situations where (P) has a special structure, e.g., if the feasible set can be decomposed into a moderate number of convex branches while the objective function is convex. Within the assumption below, which will be standing throughout this section, we quantify the requirements regarding the subproblem solver.

ASSUMPTION 1. *In each iteration $k \in \mathbb{N}$ of Algorithm 1, the approximate solution $x^{k+1} \in C$ of (2.3) satisfies*

$$\forall x \in C: \quad L_{\rho_k}(x^{k+1}, v^k, w^k) - \varepsilon_k \le L_{\rho_k}(x, v^k, w^k), \tag{2.7}$$

where $\varepsilon_k \ge 0$ is some given constant.

Typically, the inexactness parameter ε_k in Assumption 1 is chosen to be positive. While $\varepsilon_k := 0$ corresponds to the situation where the subproblems (2.3) are solved exactly, we will see that the augmented Lagrangian technique generally works fine if only approximate solutions of the subproblems are computed. This also has the advantage that whenever $\inf_x\{L_{\rho_k}(x, v^k, w^k) \,|\, x \in C\}$ is finite, then one can always find points x^{k+1} satisfying (2.7) for arbitrarily small $\varepsilon_k > 0$, while an exact global minimizer may not exist. Furthermore, we note that, due to (2.1), $L_{\rho_k}(x^{k+1}, v^k, w^k) < \infty$ holds for each $k \in \mathbb{N}$, i.e.,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$x^{k+1} \in \operatorname{dom} f \cap \operatorname{dom} g \cap C$ is valid for each computed iterate. Finally, it is worth mentioning that the validity of (2.7) guarantees that $L_{\rho_k}(\cdot, v^k, w^k)$ is bounded from below on C.

Throughout the section, we make use of the following lemma:

LEMMA 2.4. *Let $v \in \mathbb{R}^m$, $w \in Y$, and $\rho > 0$, and let $x \in \mathcal{F}$ be an arbitrary feasible point of (P). Then $L_\rho(x, v, w) \leq f(x)$ is valid.*

Proof. Due to $h(x) = 0$ and by definition of the augmented Lagrangian function L_ρ from (2.2), we find

$$L_\rho(x, v, w) = f(x) + \frac{1}{2\rho} \sum_{i=1}^{m} \left(\max^2(0, v_i + \rho\, g_i(x)) - v_i^2 \right),$$

i.e., in order to show the claim, it is sufficient to verify $\max^2(0, v_i + \rho\, g_i(x)) \leq v_i^2$ for all $i \in \{1, \ldots, m\}$. Thus, fix $i \in \{1, \ldots, m\}$ arbitrarily. In the case $v_i + \rho\, g_i(x) \leq 0$, we find $\max^2(0, v_i + \rho\, g_i(x)) = 0 \leq v_i^2$. Conversely, $v_i + \rho\, g_i(x) > 0$ yields $0 \leq v_i + \rho\, g_i(x) \leq v_i$ since $g_i(x) \leq 0$ is valid by the feasibility of x for (P), so by monotonicity of the square on the nonnegative real line, $\max^2(0, v_i + \rho\, g_i(x)) \leq v_i^2$ follows. □

Let us now start with the convergence analysis associated with Algorithm 1. Therefore, we first study issues related to the feasibility of accumulation points.

PROPOSITION 2.5. *Let (2.1) hold. Assume that Algorithm 1 produces a sequence $\{x^k\}$ such that Assumption 1 holds for some bounded sequence $\{\varepsilon_k\}$, and let $\{\rho_k\}$ and $\{v^k\}$ be the associated sequences of penalty parameters and Lagrange multiplier estimates associated with the inequality constraints in (P), respectively. Let the subsequence $\{x^{k+1}\}_{k \in K}$ and $\bar{x} \in X$ be chosen such that $x^{k+1} \rightharpoonup_K \bar{x}$. Then $\bar{x}$ is a global minimizer of the optimization problem*

$$\min_{x \in X} \ \frac{1}{2} \left\| \max(g(x), 0) \right\|_2^2 + \frac{1}{2} \left\| h(x) \right\|_Y^2 \qquad \text{s.t.} \quad x \in \operatorname{dom} f \cap C, \tag{2.8}$$

where, for each $x \notin \operatorname{dom} g$, $\frac{1}{2} \left\| \max(g(x), 0) \right\|_2^2 := \infty$. Furthermore, whenever $\mathcal{F} \neq \emptyset$, $\bar{x} \in \operatorname{dom} f \cap \mathcal{F}$ holds, and we have $V_{\rho_k}(x^{k+1}, v^k) \to_K 0$.

Proof. Let us start with the observation that, due to $x^{k+1} \in \operatorname{dom} f \cap C$ for each $k \in K$, the weak sequential lower semicontinuity of f and the weak sequential closedness of C guarantee that $\bar{x} \in \operatorname{dom} f \cap C$, i.e., $\bar{x}$ is feasible for (2.8). To proceed, we distinguish two cases.

Case 1: Suppose that $\{\rho_k\}$ remains bounded. Then Step 6 yields that ρ_k remains constant on the tail of the sequence, i.e., there is some $k_0 \in \mathbb{N}$ such that $\rho_k = \rho_{k_0}$ is valid for all $k \in \mathbb{N}$ satisfying $k \geq k_0$. In particular, condition (2.5) is satisfied for all $k \geq k_0$, which immediately yields $V_{\rho_k}(x^{k+1}, v^k) \to 0$ due to $\{x^{k+1}\} \subset \operatorname{dom} g$. On the one hand, we infer $h(x^{k+1}) \to 0$ and, on the other hand, by weak-strong sequential continuity of h, $h(x^{k+1}) \to_K h(\bar{x})$. By uniqueness of the limit, $h(\bar{x}) = 0$ follows. Due to boundedness of $\{v^k\}$, we may also assume without loss of generality that $v^k \to_K \bar{v}$ is valid for some $\bar{v} \in \mathbb{R}^m$. The componentwise weak sequential lower semicontinuity of g yields $\max(g(\bar{x}), -\bar{v}/\rho_{k_0}) \leq 0$ in the light of (2.5), i.e., $g(\bar{x}) \leq 0$ follows. Hence, $\bar{x} \in \mathcal{F}$ has been shown, i.e., $\bar{x}$ is feasible to (P). The objective value of (2.8) associated with $\bar{x}$ is 0, so that $\bar{x}$ is a global minimizer of this problem as well.

Case 2: Now, assume that $\{\rho_k\}$ is not bounded. Then by construction we have $\rho_k \to \infty$. Fix an arbitrary point $x \in \operatorname{dom} f \cap \operatorname{dom} g \cap C$ (such a point exists due to our standing assumption (2.1)). Observe that Assumption 1 and the definition of the augmented Lagrangian

ETNA
Kent State University and
Johann Radon Institute (RICAM)

function give

$$\begin{aligned} &f(x^{k+1}) + \frac{1}{2\rho_k} \sum_{i=1}^{m} \left(\max^2 \left(0, v_i^k + \rho_k g_i(x^{k+1})\right) - (v_i^k)^2 \right) \\ &\qquad + (w^k, h(x^{k+1}))_Y + \frac{\rho_k}{2} \|h(x^{k+1})\|_Y^2 - \varepsilon_k \\ &\leq f(x) + \frac{1}{2\rho_k} \sum_{i=1}^{m} \left(\max^2 \left(0, v_i^k + \rho_k g_i(x)\right) - (v_i^k)^2 \right) \\ &\qquad + (w^k, h(x))_Y + \frac{\rho_k}{2} \|h(x)\|_Y^2 \end{aligned} \tag{2.9}$$

for each $k \in \mathbb{N}$. Division by ρ_k yields

$$\begin{aligned} &\frac{f(x^{k+1})}{\rho_k} + \frac{1}{2} \sum_{i=1}^{m} \left(\max^2 \left(0, \frac{v_i^k}{\rho_k} + g_i(x^{k+1})\right) - \left(\frac{v_i^k}{\rho_k}\right)^2 \right) \\ &\qquad + (w^k/\rho_k, h(x^{k+1}))_Y + \frac{1}{2} \|h(x^{k+1})\|_Y^2 - \frac{\varepsilon_k}{\rho_k} \\ &\leq \frac{f(x)}{\rho_k} + \frac{1}{2} \sum_{i=1}^{m} \left(\max^2 \left(0, \frac{v_i^k}{\rho_k} + g_i(x)\right) - \left(\frac{v_i^k}{\rho_k}\right)^2 \right) \\ &\qquad + (w^k/\rho_k, h(x))_Y + \frac{1}{2} \|h(x)\|_Y^2 \end{aligned} \tag{2.10}$$

for each $k \in \mathbb{N}$. Note that $\{f(x^{k+1})\}_{k \in K}$ is bounded from below due to the assumed weak sequential lower semicontinuity of f. Furthermore, we have $v^k/\rho_k \to 0$, $w^k/\rho_k \to 0$, and $\varepsilon_k/\rho_k \to 0$ due to the boundedness of $\{v^k\}$, $\{w^k\}$, and $\{\varepsilon_k\}$ as well as $\rho_k \to \infty$. Thus, taking the lower limit $k \to_K \infty$ in (2.10) yields

$$\frac{1}{2} \|\max(g(\bar{x}), 0)\|_2^2 + \frac{1}{2} \|h(\bar{x})\|_Y^2 \leq \frac{1}{2} \|\max(g(x), 0)\|_2^2 + \frac{1}{2} \|h(x)\|_Y^2 < \infty, \tag{2.11}$$

where we also exploited weak sequential lower semicontinuity of $g_1, \ldots, g_m$ and weak-strong sequential continuity of h. Hence, $\bar{x}$ is a global minimizer of

$$\min_{x \in X} \ \frac{1}{2} \|\max(g(x), 0)\|_2^2 + \frac{1}{2} \|h(x)\|_Y^2 \qquad \text{s.t.} \quad x \in \operatorname{dom} f \cap \operatorname{dom} g \cap C.$$

Further, we note that for any $x \notin \operatorname{dom} g$, the objective value of this optimization problem would be ∞, so $\bar{x}$ is already a global minimizer of (2.8).

Whenever $\operatorname{dom} f \cap \mathcal{F}$ is nonempty, we may choose $x \in \operatorname{dom} f \cap \mathcal{F}$ above in order to see from (2.11) that $g(\bar{x}) \leq 0$ and $h(\bar{x}) = 0$ are valid, and $\bar{x} \in \mathcal{F}$ follows. It remains to show $V_{\rho_k}(x^{k+1}, v^k) \to_K 0$. Therefore, we exploit Lemma 2.4 and $x \in \mathcal{F}$ to obtain the estimate

$$\begin{aligned} &f(x^{k+1}) - \frac{1}{2\rho_k} \|v^k\|_2^2 + (w^k, h(x^{k+1}))_Y - \varepsilon_k \\ &\leq f(x^{k+1}) + \frac{1}{2\rho_k} \sum_{i=1}^{m} \left(\max^2 \left(0, v_i^k + \rho_k g_i(x^{k+1})\right) - (v_i^k)^2 \right) \\ &\qquad + (w^k, h(x^{k+1}))_Y + \frac{\rho_k}{2} \|h(x^{k+1})\|_Y^2 - \varepsilon_k \\ &\leq f(x) \end{aligned} \tag{2.12}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

for each $k \in \mathbb{N}$ from (2.9). From $x^{k+1} \rightharpoonup_K \bar{x}$, we find $h(x^{k+1}) \to_K h(\bar{x})$ by weak-strong sequential continuity of h. Thus, $\{(w^k, h(x^{k+1}))_Y\}_{k\in K}$ remains bounded as $\{w^k\}$ is bounded by construction. Since $\{v^k\}$ is bounded by construction as well while $\rho_k \to \infty$ holds and since $\{\varepsilon_k\}$ is assumed to be bounded, the sequence $\{f(x^{k+1})\}_{k\in K}$ is bounded from above. Its lower boundedness has already been mentioned earlier in the proof. Hence, we have $f(x^{k+1})/\rho_k \to_K 0$.

Dividing (2.12) by $\rho_k/2$ and taking, in contrast to our earlier strategy, the upper limit $k \to_K \infty$ in the second estimate gives

$$\limsup_{k\to_K\infty} \left(\| \max(0, v^k/\rho_k + g(x^{k+1}))\|_2^2 + \|h(x^{k+1})\|_Y^2 \right) \leq 0$$

as we now also know $f(x^{k+1})/\rho_k \to_K 0$. Due to $v^k/\rho_k \to 0$, this gives convergence of $\| \max(g(x^{k+1}), -v^k/\rho_k)\|_2 \to_K 0$, and it also follows that $\|h(x^{k+1})\|_Y \to_K 0$. Since all norms in finite-dimensional spaces are equivalent, $V_{\rho_k}(x^{k+1}, v^k) \to_K 0$ is obtained, and the proof is completed. □

Next, we want to show that under Assumption 1, Algorithm 1 can be used to compute a global minimizer of (P) provided there exists one.

THEOREM 2.6. *Let* $\operatorname{dom} f \cap \mathcal{F} \neq \emptyset$ *hold. Assume that Algorithm 1 produces a sequence* $\{x^k\}$ *such that Assumption 1 holds for some sequence* $\{\varepsilon_k\}$ *satisfying* $\varepsilon_k \to 0$. *Then, for each subsequence* $\{x^{k+1}\}_{k\in K}$ *and each point* $\bar{x} \in X$ *satisfying* $x^{k+1} \rightharpoonup_K \bar{x}$, *we have* $f(x^{k+1}) \to_K f(\bar{x})$, *and* $\bar{x}$ *is a global minimizer of* (P).

Proof. To start, note that Proposition 2.5 guarantees that $\bar{x}$ is a feasible point of (P). Furthermore, for each feasible point $x \in \operatorname{dom} f \cap \mathcal{F}$ of (P), Assumption 1 and Lemma 2.4 yield

$$\forall k \in \mathbb{N}: \quad L_{\rho_k}(x^{k+1}, v^k, w^k) - \varepsilon_k \leq L_{\rho_k}(x, v^k, w^k) \leq f(x). \tag{2.13}$$

We note that the same inequality holds trivially for all $x \in \mathcal{F} \setminus \operatorname{dom} f$. We will first prove that $\limsup_{k\to_K\infty} f(x^{k+1}) \leq f(\bar{x})$ is valid. Again, we proceed by investigating two disjoint cases.

Case 1: Suppose that $\{\rho_k\}$ remains bounded. As in the proof of Proposition 2.5, this implies that condition (2.5) holds along the tail of the sequence. Thus, for each $i \in \{1, \dots, m\}$, we find

$$\left|\max\left(0, v_i^k/\rho_k + g_i(x^{k+1})\right) - v_i^k/\rho_k\right| = \left|\max\left(g_i(x^{k+1}), -v_i^k/\rho_k\right)\right| \to 0$$

as $k \to \infty$. By the boundedness of $\{v_i^k/\rho_k\}$, the term $\{\max(0, v_i^k/\rho_k + g_i(x^{k+1}))\}$ needs to be bounded as well, which is why we already find that

$$\left|\max^2\left(0, v_i^k/\rho_k + g_i(x^{k+1})\right) - \left(v_i^k/\rho_k\right)^2\right| \to 0,$$

and by the boundedness of $\{\rho_k\}$ this yields

$$\frac{1}{\rho_k}\left(\max^2(0, v_i^k + \rho_k\, g_i(x^{k+1})) - (v_i^k)^2\right) \to 0.$$

Furthermore, we find $(w^k, h(x^{k+1}))_Y \to_K 0$ and $\frac{\rho_k}{2}\|h(x^{k+1})\|_Y^2 \to_K 0$ from the fact that $x^{k+1} \rightharpoonup_K \bar{x}$, the weak-strong sequential continuity of h, that $h(\bar{x}) = 0$, and the boundedness of $\{w^k\}$. Plugging all this into (2.13) while respecting the definition of the function L_{ρ_k} and $\varepsilon_k \to 0$, we find $\limsup_{k\to_K\infty} f(x^{k+1}) \leq f(x)$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Case 2: Let $\{\rho_k\}$ be unbounded. Then we already have $\rho_k \to \infty$ by construction of Algorithm 1. Furthermore, (2.13) implies the validity of the estimate

$$\forall k \in \mathbb{N}: \quad f(x^{k+1}) - \frac{1}{2\rho_k}\|v^k\|_2^2 + (w^k, h(x^{k+1}))_Y - \varepsilon_k \leq f(x)$$

by leaving out some of the nonnegative terms on the left-hand side. As above, we find $(w^k, h(x^{k+1})) \to_K 0$ by the boundedness of $\{w^k\}$, the weak-strong sequential continuity of h, and by $h(\bar{x}) = 0$. The boundedness of $\{v^k\}$ and $\rho_k \to \infty$ yield $\frac{1}{2\rho_k}\|v^k\|_2^2 \to 0$ as $k \to \infty$. Thus, taking the upper limit in the above estimate shows $\limsup_{k\to_K\infty} f(x^{k+1}) \leq f(x)$.

In order to finalize the proof, we observe that the weak sequential lower semicontinuity of f now yields the estimate

$$f(\bar{x}) \leq \liminf_{k\to_K\infty} f(x^{k+1}) \leq \limsup_{k\to_K\infty} f(x^{k+1}) \leq f(x).$$

As this has been shown for each $x \in \operatorname{dom} f \cap \mathcal{F}$ (and it is trivially valid for each $x \in \mathcal{F}\backslash\operatorname{dom} f$), $\bar{x}$ is a global minimizer of (P). Using the above estimate with $x := \bar{x}$, we additionally find the convergence $f(x^{k+1}) \to_K f(\bar{x})$. □

As a consequence of the previous result, we obtain the following stronger version for convex problems with a strongly convex objective function.

COROLLARY 2.7. *Let $\mathcal{F} \neq \emptyset$ hold. Assume that Algorithm 1 produces a sequence $\{x^k\}$ such that Assumption 1 holds for some sequence $\{\varepsilon_k\}$ satisfying $\varepsilon_k \to 0$. Furthermore, let f be continuous as well as strongly convex, $g_1, \ldots, g_m$ be convex, h be affine, and C be convex. Then the entire sequence $\{x^k\}$ converges (strongly) to the uniquely determined global minimizer of* (P).

Proof. Since f is strongly convex, the (convex) optimization problem (P) has a unique solution $\bar{x} \in X$; see [53, Theorem 2.5.1, Propositions 2.5.6, 3.5.8]. As $\bar{x}$ is a minimizer of the underlying convex problem (P) and since f is assumed to be continuous, there exists $\bar{\xi} \in \partial f(\bar{x})$ such that $\langle \bar{\xi}, x - \bar{x}\rangle_X \geq 0$ is valid for all $x \in \mathcal{F}$; see [53, Theorem 2.9.1]. By strong convexity of f, there exists a constant $\nu > 0$ such that

$$\forall x \in X: \quad f(x) \geq f(\bar{x}) + \langle \bar{\xi}, x - \bar{x}\rangle_X + \frac{\nu}{2}\|x - \bar{x}\|_X^2; \tag{2.14}$$

see, e.g., [53, Corollary 3.5.11]. This implies

$$\begin{aligned}
f(\bar{x}) + \langle \bar{\xi}, x^{k+1} - \bar{x}\rangle_X + \frac{\nu}{2}\|x^{k+1} - \bar{x}\|_X^2 - \frac{1}{2\rho_k}\|v^k\|_2^2 + (w^k, h(x^{k+1}))_Y \\
&\leq f(x^{k+1}) - \frac{1}{2\rho_k}\|v^k\|_2^2 + (w^k, h(x^{k+1}))_Y \\
&\leq f(x^{k+1}) + \frac{1}{2\rho_k}\sum_{i=1}^{m}\left(\max^2\left(0, v_i^k + \rho_k g_i(x^{k+1})\right) - (v_i^k)^2\right) \\
&\qquad + (w^k, h(x^{k+1}))_Y + \frac{\rho_k}{2}\|h(x^{k+1})\|_Y^2 \\
&= L_{\rho_k}(x^{k+1}, v^k, w^k) \\
&\leq L_{\rho_k}(\bar{x}, v^k, w^k) + \varepsilon_k \\
&\leq f(\bar{x}) + \varepsilon_k
\end{aligned}$$

for all $k \in \mathbb{N}$, where the first inequality results from (2.14), the second one comes from adding some nonnegative terms, the subsequent equation is simply the definition of the augmented

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Lagrangian, the penultimate inequality takes into account Assumption 1, and the final estimate uses Lemma 2.4.

Note that, on the one hand, the term on the right-hand side is bounded. On the other hand, since $\{v^k\}$ and $\{w^k\}$ are bounded sequences and h is affine, the growth behavior of the left-hand side is dominated by the quadratic term. Consequently, the sequence $\{x^k\}$ is bounded and therefore has a weakly convergent subsequence in the reflexive space X. The weak limit is necessarily a solution of (P) by Theorem 2.6. Since the entire sequence $\{x^k\}$ is bounded, we therefore get $x^k \rightharpoonup \bar{x}$ and $f(x^k) \to f(\bar{x})$ from Theorem 2.6.

Let us now test (2.14) with $x := x^{k+1}$. Then, after some rearrangements, we find

$$f(x^{k+1}) - f(\bar{x}) - \langle \bar{\xi}, x^{k+1} - \bar{x} \rangle_X \geq \frac{\nu}{2} \|x^{k+1} - \bar{x}\|_X^2 \geq 0.$$

From $x^{k+1} \rightharpoonup \bar{x}$ and $f(x^{k+1}) \to f(\bar{x})$, the left-hand side in this estimate tends to 0 as $k \to \infty$. Due to $\nu > 0$, this immediately gives $x^{k+1} \to \bar{x}$, and the proof is complete. ∎

We end this section by discussing a suitable termination criterion for Algorithm 1 and recalling the central convergence result for convex problems stated in [44], which is based on a non-safeguarded version of Algorithm 1.

REMARK 2.8. Observe that Proposition 2.5 indicates that testing the condition $V_{\rho_{k-1}}(x^k, v^{k-1}) \leq \varepsilon_{\text{abs}}^{\text{alm}}$ for some $\varepsilon_{\text{abs}}^{\text{alm}} \geq 0$ in each of the iterations $k \in \mathbb{N}$, $k \geq 1$, is a reasonable termination criterion for Algorithm 1 provided that $\operatorname{dom} f \cap \mathcal{F} \neq \emptyset$. On the one hand, if $V_{\rho_{k-1}}(x^k, v^{k-1})$ is small, then the underlying point x^k is close to be feasible, and the associated Lagrange multiplier estimate v^{k-1} is close to satisfy the associated complementarity-slackness condition with respect to the inequality constraints. On the other hand, along weakly convergent subsequences of the iterates produced by Algorithm 1, $V_{\rho_{k-1}}(x^k, v^{k-1})$ indeed becomes arbitrarily small under the assumptions of Proposition 2.5. Furthermore, under the assumptions of Theorem 2.6, weak accumulation points are already global minimizers of (P).

REMARK 2.9. In [44], the author considers (P) with convex, continuously differentiable functions $f, g_1, \ldots, g_m$, a closed, convex set C, and in the absence of (infinite-dimensional) equality constraints. For the numerical solution of this problem, a simplified version of Algorithm 1 is suggested where safeguarding of the Lagrange multipliers is omitted and the penalty parameter stays constant. More precisely, this means that Steps 3 and 6 are removed from the method, in Step 4, the subproblem

$$\min_{x \in X} L_{\rho_0}(x, \lambda^k) \qquad \text{s.t.} \quad x \in C \tag{2.15}$$

is solved up to ε_k-minimality (see Assumption 1), and the Lagrange multiplier update in Step 5 is replaced by

$$\lambda^{k+1} := \max\bigl(0, \lambda^k + \rho_0 g(x^{k+1})\bigr).$$

In [44, Theorem 2.1], it is shown that whenever (P) possesses a minimizer which is stationary (in the sense that Lagrange multipliers exist) while $\sum_{k=0}^{\infty} \sqrt{\varepsilon_k} < \infty$ holds, then each accumulation point $(\bar{x}, \bar{\lambda}) \in X \times \mathbb{R}_+^m$ of the primal-dual sequence $\{(x^k, \lambda^k)\}$ produced by the adjusted method is a Karush–Kuhn–Tucker pair of (P), i.e., $\bar{x}$ is a feasible point with associated Lagrange multipliers $\bar{\lambda}$. In particular, $\bar{x}$ is a minimizer of (P) in this case.

3. Numerical solution of the Poisson denoising model. In this section, we numerically test the variational Poisson denoising model. First, in Section 3.1, we introduce an affine reformulation of the constraints in (VPD). General comments about the implementation of the experiments are presented in Section 3.2. In particular, we introduce three augmented Lagrangian schemes which are used to tackle (VPD) computationally, one of them focusing

ETNA
Kent State University and
Johann Radon Institute (RICAM)

on the original model (VPD) while the other two address the aforementioned reformulation. In Section 3.3, we comment on a stochastic gradient descent method which is used to solve the arising subproblems within the augmented Lagrangian schemes. Our way of documenting the obtained results is briefly outlined in Section 3.4. Finally, the outcome of our experiments is presented in Section 3.5.

3.1. Reformulation of the constraints. Let us point out that the constraints in (VPD) are heavily nonlinear and, due to the fact that the Kullback–Leibler divergence is extended real-valued and nonsmooth. In this section, we show that these constraints can be reformulated as affine ones, making the associated reformulation of (VPD) seemingly easier to tackle from an algorithmic point of view.

For given $a \geq 0$ and $c > 0$, let us consider the nonlinear constraint

$$\eta(a, b) \leq c \tag{3.1}$$

for the real variable b. Above, η is the Kullback–Leibler divergence defined in (1.7). Whenever $a = 0$, we can immediately rewrite (3.1) as $0 \leq b \leq c$.

Thus, let us assume that $a > 0$. Then (3.1) requires $b > 0$. Let us study the properties of the smooth function $(0, \infty) \ni b \mapsto \eta(a, b) \in \mathbb{R}$. One can easily verify that it is strictly convex with a uniquely determined minimizer at $\check{b} := a$ with function value 0. Furthermore, we have

$$\lim_{b \downarrow 0} \eta(a, b) = \infty, \qquad \lim_{b \to \infty} \eta(a, b) = \infty.$$

Hence, there exist uniquely determined points $\underline{b} \in (0, \check{b})$ and $\overline{b} \in (\check{b}, \infty)$ that solve the nonlinear equation

$$\eta(a, b) - c = 0, \tag{3.2}$$

and thus, (3.1) is equivalent to $\underline{b} \leq b \leq \overline{b}$. Let us now describe how the bounds $\underline{b}$ and $\overline{b}$ can be accessed computationally. Division by a in (3.2) and some rearrangements yield

$$\frac{b}{a} + \ln\left(\frac{a}{b}\right) = \frac{c}{a} + 1.$$

We exponentiate this equation and take the reciprocal on both sides in order to find

$$\left(-\frac{b}{a}\right) \exp\left(-\frac{b}{a}\right) = \frac{-1}{\exp\left(\frac{c}{a} + 1\right)}.$$

Due to

$$\frac{-1}{\exp(1)} < \frac{-1}{\exp\left(\frac{c}{a} + 1\right)} < 0,$$

the latter equation indeed possesses two solutions, which can be expressed in terms of *Lambert's W function* W_κ, i.e., the multivalued inverse of $t \mapsto t \exp(t)$ (see, e.g., [16] for properties, applications, and the numerical evaluation of this function), as

$$\underline{b} := \underline{b}(a, c) := -aW_0\left(\frac{-1}{\exp\left(\frac{c}{a} + 1\right)}\right), \qquad \overline{b} := \overline{b}(a, c) := -aW_{-1}\left(\frac{-1}{\exp\left(\frac{c}{a} + 1\right)}\right).$$

Taking all our above findings together, we can equivalently state (VPD) as

$$(\text{VPD}_{\text{aff}}) \qquad \min_{u \in L^2(\Omega)} f(u) \qquad \text{s.t.} \quad \begin{cases} 0 \leq u_B \leq r(|B|) & \forall B \in \mathcal{B}_Z^0, \\ \underline{b}(Z_B, r(|B|)) \leq u_B \leq \overline{b}(Z_B, r(|B|)) & \forall B \in \mathcal{B}_Z^+, \end{cases}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

where we made use of the sets

$$\mathcal{B}_Z^0 := \{B \in \mathcal{B} \,|\, Z_B = 0\}, \qquad \mathcal{B}_Z^+ := \{B \in \mathcal{B} \,|\, Z_B > 0\}.$$

Observe that (VPD$_{\text{aff}}$) possesses twice as many constraints as (VPD) does.

Let us note that (VPD$_{\text{aff}}$) is *not* a box-constrained problem since u_B already involves an averaging operation of u on the box $B \in \mathcal{B}$. However, using the weighted characteristic function $\overline{\chi}_B \in L^2(\Omega)$ given by

$$\forall \omega \in \Omega\colon \quad \overline{\chi}_B(\omega) := \begin{cases} |B|^{-1} & \omega \in B, \\ 0 & \omega \notin B, \end{cases}$$

for each subbox $B \in \mathcal{B}$, the constraints in (VPD$_{\text{aff}}$) can be rewritten in the form

$$\begin{cases} 0 \le \langle \overline{\chi}_B, u \rangle_{L^2(\Omega)} \le r(|B|) & \forall B \in \mathcal{B}_Z^0, \\ \underline{b}(Z_B, r(|B|)) \le \langle \overline{\chi}_B, u \rangle_{L^2(\Omega)} \le \overline{b}(Z_B, r(|B|)) & \forall B \in \mathcal{B}_Z^+. \end{cases}$$

Hence, whenever the objective function f is smooth enough, local minimizers of (VPD$_{\text{aff}}$) are stationary point of that problem due to [11, Proposition 2.42], i.e., Lagrange multipliers exist in that case.

3.2. Implementation. For the numerical realization, we discretize the image using 256^2 equally sized pixels. We use a zero padding of five equally sized pixels at the borders of the image to prevent artifacts in the reconstruction due to the periodic extension in the objective value associated with (1.4). Thus, setting $n := 266$, the image to be denoised has actually a size of n^2 pixels. A wider padding would give a higher guarantee of avoiding these artifacts at the cost of higher computational complexity. Empirically, our tests show that the padding size of five pixels seems to be sufficient. The (padded) image u is therefore approximated by an $n \times n$ matrix of pixels with pixel size $s := 1/n^2 = 266^{-2}$. With this resolution, the family of regions $\mathcal{B} \subset 2^\Omega$ is chosen as all subsquares of the image with side length (scale) between 1 and 64 pixels. The size $|B|$ of a region $B \in \mathcal{B}$ is numerically computed as $|B| := s\,\#B$. The number of constraints in (VPD), i.e., the number of subboxes in $\mathcal{B}$, is then calculated by

$$\sum_{i=n-63}^{n} i^2 = 3\,541\,216$$

and is approximately 50 times higher than the total number of pixels. The usage of squares is clearly subjective, but without a priori information about the image content, it is not clear which other shapes might be preferable. We emphasize that, however, our methodology is also applicable with other shapes such as rectangles, circles, or ellipses.

As there are way more subsquares with small side length, a penalty term

$$\operatorname{pen}(|B|) := \sqrt{2\,(\log\,(n^2/\,|B|) + 1)},$$

which only depends on the size of the subsquares, is introduced. This is necessary to avoid the small subsquares to dominate the statistical behavior of the overall test statistic

$$T_n(Z, u, \mathcal{B}) := \max_{B \in \mathcal{B}} [T_B(Z, u) - \operatorname{pen}(|B|)]$$

(see [38]), where T_B is the LRT statistic from (1.6). We approximate the $(1-\alpha)$-quantile $q_{1-\alpha}$ of T_n by the (empirically sampled) $(1-\alpha)$-quantile $\tilde{q}_{1-\alpha}$ of

$$M_n(\mathcal{B}) := \max_{B \in \mathcal{B}} \left[|B|^{-1/2} \left| \sum\nolimits_{i \in B} X_i \right| - \operatorname{pen}(|B|) \right]$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

with i.i.d. standard normal random variables X_i. If the smallest scale in $\mathcal{B}$ is at least of size $\log(n)$, then this approximation is shown to be valid in [38]. However, the chosen penalization pen effectively overdamps the small scales (see [46]), which makes this approximation reasonable over all scales considered here. Altogether, this leads to the right-hand side

$$r(|B|) := \frac{(\tilde{q}_{1-\alpha} + \mathrm{pen}(|B|))^2}{2|B|}$$

in (VPD). In the numerical experiments, the 0.1-quantile $\tilde{q}_{0.1} := 1.63$ is used because for larger values of α, the local hypothesis tests are not restrictive enough such that the obtained reconstruction is oversmoothed. For the same reason, $\theta := 0.01$ is chosen relatively small in the Sobolev-type penalty (1.4), which is used as the objective function subsequently. Observe that the discretized problem associated with ($\mathrm{VPD}_{\mathrm{aff}}$) is a convex quadratic optimization problem in this situation.

We solve the associated problem (VPD) with the aid of the following procedures:

- **ALM**: the safeguarded augmented Lagrangian method from Algorithm 1 applied to (VPD),
- **ALMr**: the safeguarded augmented Lagrangian method from Algorithm 1 applied to the reformulated problem ($\mathrm{VPD}_{\mathrm{aff}}$), and
- **sALMr**: the simplified augmented Lagrangian method without safeguarding and with constant penalty parameter applied to ($\mathrm{VPD}_{\mathrm{aff}}$); see Remark 2.9.

For all three methods, the discretized noisy observation is taken as the primal starting point, i.e., $x^0 := Z$, while $\lambda^0 := 0$ is chosen for the initial Lagrange multiplier. Note that this choice for x^0 is only possible in the discretized setting, as Z is likely to lack L^2-regularity in the infinite-dimensional framework. In the case of continuous computations, one could, e.g., use a kernel density estimator to obtain some point $x^0 \in L^2(\Omega)$ from Z. Within our implementation of **ALM** and **ALMr**, we choose v^k as the componentwise projection of the Lagrange multiplier λ^k onto the interval $[0, 10^8]$. Furthermore, we use the parameters $\rho_0 := 4$, $\tau := 0.9$, and $\gamma := 4$ for these methods. In **sALMr**, we make use of the (constant) penalty parameter $\rho_0 := 4 \cdot 10^5$. Finally, for the abort of all three methods, we exploit the termination criterion from Remark 2.8 with $\varepsilon_{\mathrm{abs}}^{\mathrm{alm}} := 10^{-2}$. Let us note that the aforementioned choice for ρ_0 in **sALMr** is such that this termination criterion is almost satisfied already in the first iteration of the method. In some preliminary experiments, it turned out that **sALMr** is not capable to enhance the feasibility properties along the iterates significantly if ρ_0 is chosen much smaller than $4 \cdot 10^5$. Finally, all three methods are, at the latest, aborted after a total number of 30 outer augmented Lagrangian iterations.

3.3. Stochastic gradient descent as a subproblem solver. Solving the unconstrained associated subproblems (2.3) as well as (2.15) within the algorithmic frameworks **ALM** as well as **ALMr** and **sALMr**, respectively, is computationally expensive, especially as the problems (VPD) and ($\mathrm{VPD}_{\mathrm{aff}}$) possess many constraints. This obstacle is tackled by using the first-order gradient descent method NADAM from [23], which outperformed other gradient descent methods in the setting we are considering here. It is also utilized that the constraints are redundant to a certain degree, and thus, a stochastic version of the NADAM method can be used. Undoubtedly, this also speeds up the evaluation of the (stochastic) gradient of the augmented Lagrangian function and cheapens each iteration of the subproblem solver.

Let us comment on the implementation of NADAM in the context of **ALM**. For the fixed penalty parameter $\rho_k > 0$ and the Lagrange multiplier estimate $v^k := \{v_B^k\}_{B \in \mathcal{B}}$, the augmented Lagrangian subproblem (2.3) takes the particular form

$$\min_{u \in \mathbb{R}^{n \times n}} f(u) + \frac{1}{2\rho_k} \sum_{B \in \mathcal{B}} \left(\max^2\left(0, v_B^k + \rho_k(\eta(Z_B, u_B) - r(|B|))\right) - (v_B^k)^2 \right)$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

in the present situation. Here and in what follows, we approximate the continuous mean $|B|^{-1}\int_B u(\omega)\,\mathrm{d}\omega$ by $u_B := s|B|^{-1}\sum_{i\in B} u_i = (\#B)^{-1}\sum_{i\in B} u_i$, which corresponds to the discrete mean. For the NADAM method, one needs to calculate the gradient of the augmented Lagrangian function. Therefore, the partial derivative of $u \mapsto \eta(Z_B, u_B)$ with respect to the pixel u_i (where it exists) is given by $(\#B)^{-1}(1 - Z_B/u_B)$ if $i \in B$ and 0 otherwise. Thus, the partial derivative of the associated augmented Lagrangian function with respect to the pixel u_i (where it exists) equals

$$f'_{u_i}(u) + \sum_{B\in\mathcal{B}(i)} \frac{1}{\#B} \max\left(0, v_B^k + \rho_k(\eta(Z_B, u_B) - r(|B|))\right)\left(1 - \frac{Z_B}{u_B}\right),$$

where

$$\mathcal{B}(i) := \{B \in \mathcal{B} \,|\, i \in B\}.$$

The above formula is valid whenever $u_B > 0$ for all $B \in \mathcal{B}(i)$. To account for the non-differentiability on the boundary, we set

$$(L_{\rho_k})'_{u_i}(u, v^k) := f'_{u_i}(u) + \sum_{B\in\mathcal{B}(i)} b_{\rho_k}(Z, u, B, v^k)$$

with

$$b_{\rho_k}(Z, u, B, v^k) := \begin{cases} \frac{1}{\#B} \max\left(0, v_B^k + \rho_k(\eta(Z_B, u_B) - r(|B|))\right)\left(1 - \frac{Z_B}{u_B}\right) & \text{if } u_B > 0, \\ C & \text{if } u_B = 0 \text{ and } Z_B > 0, \\ 0 & \text{if } u_B = Z_B = 0. \end{cases}$$

In the case $u_B = Z_B = 0$, the constraint is satisfied, and thus, we can set the corresponding gradient to 0. The rationale behind the definition for $u_B = 0$ and $Z_B > 0$ is to enforce a step in positive direction. Numerically, we use $C := -10 < 0$. By definition it always holds that $Z_B \geq 0$, and therefore we do not need to cope with the case $Z_B < 0$. As the NADAM method may produce iterates having pixels with negative value, we set all pixels to zero which have negative value after each NADAM iteration. This way, we also ensure $u_B \geq 0$.

Instead of calculating the summand for every $B \in \mathcal{B}$, we choose a random family $\mathcal{B}_r \subset \mathcal{B}$ and approximate the gradient by

$$(L_{\rho_k})'_{u_i}(u, v^k) \approx f'_{u_i}(u) + \sum_{B\in\mathcal{B}_r\cap\mathcal{B}(i)} b_{\rho_k}(Z, u, B, v^k).$$

As it is possible to efficiently calculate all summands with the same scale $|B|$ with the help of the discrete Fourier transform, we pick the scales at random and only include all sets B of those scales in $\mathcal{B}_r$. In practice, it was first tried to use a fixed number of 10 scales. This yielded fast convergence in the beginning, but convergence slowed down during the runs due to missing accuracy when solving the subproblems. Thus, we decided to increase the number of scales picked during the algorithm although this worsens the running time of a single augmented Lagrangian step. More precisely, in our experiments, we now increase the amount of scales by one after every augmented Lagrangian step. For simplicity, a fixed number of 300 iterations of the NADAM method is chosen, and the stepsize is picked constant as $\max(0.005, 0.8^k)$ in the k-th iteration of Algorithm 1 to solve the augmented Lagrangian subproblem (2.3).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Within the algorithmic framework of **ALMr**, for a given penalty parameter $\rho_k > 0$ and Lagrange multiplier estimates $v^k := \{\underline{v}_B^k\}_{B\in\mathcal{B}} \cup \{\overline{v}_B^k\}_{B\in\mathcal{B}}$, the associated augmented Lagrangian subproblem (2.3) is given by

$$\begin{aligned}\min_{u\in\mathbb{R}^{n\times n}} f(u) + \frac{1}{2\rho_k} \sum_{B\in\mathcal{B}_Z^0} \Big(&\max^2\big(0, \overline{v}_B^k + \rho_k(u_B - r(|B|))\big) - (\overline{v}_B^k)^2 \\ &+ \max^2\big(0, \underline{v}_B^k - \rho_k u_B\big) - (\underline{v}_B^k)^2\Big) \\ + \frac{1}{2\rho_k} \sum_{B\in\mathcal{B}_Z^+} \Big(&\max^2\big(0, \overline{v}_B^k + \rho_k(u_B - \overline{b}(Z_B, r(|B|)))\big) - (\overline{v}_B^k)^2 \\ &+ \max^2\big(0, \underline{v}_B^k + \rho_k(\underline{b}(Z_B, r(|B|)) - u_B)\big) - (\underline{v}_B^k)^2\Big).\end{aligned} \tag{3.3}$$

In the discretized setting, the partial derivative of the augmented Lagrangian function with respect to the pixel u_i equals

$$\begin{aligned}f'_{u_i}(u) + \sum_{B\in\mathcal{B}_Z^0\cap\mathcal{B}(i)} \frac{1}{\#B}\Big(&\max\big(0, \overline{v}_B^k + \rho_k(u_B - r(|B|))\big) - \max\big(0, \underline{v}_B^k - \rho_k u_B\big)\Big) \\ + \sum_{B\in\mathcal{B}_Z^+\cap\mathcal{B}(i)} \frac{1}{\#B}\Big(&\max\big(0, \overline{v}_B^k + \rho_k(u_B - \overline{b}(Z_B, r(|B|)))\big) \\ &- \max\big(0, \underline{v}_B^k + \rho_k(\underline{b}(Z_B, r(|B|)) - u_B)\big)\Big).\end{aligned}$$

Again, we do not exploit the full gradient of the augmented Lagrangian function in the NADAM framework but rely on an approximation where the appearing sums are restricted to a random family $\mathcal{B}_r \subset \mathcal{B}$ which is chosen in the same way as outlined above. Similarly, the maximum number of iterations and the stepsize for NADAM are chosen as described in the setting of **ALM**.

Finally, in the setting of **sALMr**, the appearing augmented Lagrangian subproblem (2.15) equals (3.3) with $\rho_k := \rho_0$ and $v^k := \lambda^k$ for the current Lagrange multiplier $\lambda^k := \{\underline{\lambda}_B^k\}_{B\in\mathcal{B}} \cup \{\overline{\lambda}_B^k\}_{B\in\mathcal{B}}$, and this subproblem is treated in the same way as (3.3) in the **ALMr** framework. The only exception is that we choose the number of NADAM iterations by 10^4 in the first outer iteration of the augmented Lagrangian method. This is necessary to get a good approximate solution of the subproblem. In the subsequent outer iterations, we again use merely 300 NADAM iterations as we are already closer to a good approximate solution.

3.4. Documentation of results. In order to compare the algorithmic frameworks in $\mathcal{A} := \{\textbf{ALM}, \textbf{ALMr}, \textbf{sALMr}\}$ in a reasonable way, we run all these methods on the same set $\mathcal{P}$ of benchmark instances for (VPD) and illustrate the outcome with the aid of so-called performance profiles (see [22]) based on different performance indicators.

Let us briefly comment on the concept of performance profiles. Therefore, we denote the output of algorithm $\mathrm{a} \in \mathcal{A}$ for problem $\mathrm{p} \in \mathcal{P}$ by $u_{\mathrm{p}}^{\mathrm{a}}$. Let q be a given performance measure taking only positive values such that a smaller value of q indicates a better performance. The associated performance ratio is defined via

$$\forall \mathrm{p}\in\mathcal{P},\ \forall \mathrm{a}\in\mathcal{A}: \quad r_{\mathrm{p,a}} := \frac{q(u_{\mathrm{p}}^{\mathrm{a}})}{\min\{q(u_{\mathrm{p}}^{\mathrm{a}'}) \,|\, \mathrm{a}'\in\mathcal{A}\}}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

In the associated performance profile, for each of the algorithms $\mathrm{a} \in \mathcal{A}$, we plot the illustrative part of the nondecreasing curve $\zeta_{\mathrm{a}} \colon [1, \infty) \to [0, 1]$ given by

$$\forall t \in [1, \infty) \colon \quad \zeta_{\mathrm{a}}(t) := \frac{\#\{\mathrm{p} \in \mathcal{P} \,|\, r_{\mathrm{p,a}} \leq t\}}{\#\mathcal{P}}.$$

For example, for $\mathrm{a} \in \mathcal{A}$, $\zeta_{\mathrm{a}}(1)$ denotes the portion of benchmark problems in $\mathcal{P}$ on which algorithm a performed best. For $t > 1$, $\zeta_{\mathrm{a}}(t)$ represents the portion of benchmark problems $\mathrm{p} \in \mathcal{P}$ on which algorithm a produces an output whose value with respect to the performance measure $q(u_{\mathrm{p}}^{\mathrm{a}})$ is at most the t-fold of the best value achieved by any algorithm in $\mathcal{A}$ for the particular instance p.

To document the results of our experiments, we make use of the performance measures:

- *Objective function value of output*
  We note that the Sobolev-type penalty f from (1.4) only takes nonnegative values. In all practically relevant scenarios, all feasible points of (VPD) will possess positive objective values for this choice of the objective function.
- *Number of outer augmented Lagrangian iterations until termination*
- *Value of error measure when the method is aborted*
  For $\mathrm{a} \in \{\textbf{ALM}, \textbf{ALMr}\}$, the performance measure is given by $V_{\rho_{k-1}}(u_{\mathrm{p}}^{\mathrm{a}}, v^{k-1})$, where $k \in \mathbb{N}$ is the final value of the iteration counter when Algorithm 1 is aborted. For $\mathrm{a} := \textbf{sALMr}$, we take $V_{\rho_0}(u_{\mathrm{p}}^{\mathrm{a}}, \lambda^{k-1})$ as the performance measure, where the meaning of k is the same as mentioned above.
- *Value of the penalty parameter when the method is aborted*
  Here, we only compare **ALM** and **ALMr** since the penalty parameter is not enlarged in **sALMr**.
- *Percentage of constraints violated by the output*
  Let us note that a constraint in (VPD) is violated if and only if precisely one of the associated two constraints in ($\mathrm{VPD}_{\mathrm{aff}}$) is violated. In particular, an arbitrarily chosen point satisfies at least 50% of the constraints in ($\mathrm{VPD}_{\mathrm{aff}}$). We thus measure the percentage of violated constraints in terms of the constraints in (VPD) for all three methods in order to state a fair comparison. Due to the huge number of constraints, it is unlikely that any of the methods actually computes a feasible point of (VPD) or ($\mathrm{VPD}_{\mathrm{aff}}$). Nevertheless, we added the positive offset $\delta := 1$ to each of the percentages to obtain meaningful performance profiles.
- *Maximum relative (affine) constraint violation of the output*
  For a comparatively fair comparison, this quantity is first measured in terms of the constraints of (VPD) only, i.e., for any $\mathrm{p} \in \mathcal{P}$ and $\mathrm{a} \in \mathcal{A}$, the quantity

  $$\max_{B \in \mathcal{B}} \frac{\max(0, \eta(Z_B, (u_{\mathrm{p}}^{\mathrm{a}})_B) - r(|B|))}{r(|B|)}$$

  is used as a performance measure. For the outputs of **ALM**, this quantity has almost always been finite. However, as **ALMr** and **sALMr** aim to compute feasible points of ($\mathrm{VPD}_{\mathrm{aff}}$), it may happen that, for some boxes $B \in \mathcal{B}$, the appearing fraction takes the value ∞; see (1.7) again for the precise definition of the Kullback–Leibler divergence η. In this case, we set the maximum relative constraint violation to be the largest finite value of

  $$\frac{\max(0, \eta(Z_B, (u_{\mathrm{p}'}^{\mathrm{a}'})_B) - r(|B|))}{r(|B|)},$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

which is observed for all $\mathrm{p}' \in \mathcal{P}$ and $\mathrm{a}' \in \mathcal{A}$ in order to get a meaningful performance profile. Second, in a similar fashion, the relative constraint violation is measured in terms of the constraints in ($\mathrm{VPD}_{\mathrm{aff}}$) for all three methods. Note that in this case, no issues regarding infinite values have to be faced. Again, a constant offset of $\delta := 1$ is added to the relative (affine) constraint violation to make the associated performance profile compelling.

- *Average relative (affine) constraint violation of the output*
  This performance measure is first defined by

$$\sum_{B \in \mathcal{B}} \frac{\max(0, \eta(Z_B, (u_{\mathrm{p}}^{\mathrm{a}})_B) - r(|B|))}{\#B\, r(|B|)}$$

  for $\mathrm{p} \in \mathcal{P}$ and $\mathrm{a} \in \mathcal{A}$. Again, we note that the appearing fraction may take value ∞ for $\mathrm{a} \in \{\textbf{ALMr}, \textbf{sALMr}\}$. In this case, it is replaced by the largest finite value of this fraction achieved over $\mathcal{P}$ and $\mathcal{A}$. Second, a similar quantity based on the affine constraints in ($\mathrm{VPD}_{\mathrm{aff}}$) is defined for a comparison of all three methods. In both situations, we use the constant offset $\delta := 1$ to obtain meaningful performance profiles.

Let us mention that we do not rely on computation time as a performance measure for the following reasons. Whenever $k \in \mathbb{N}$ denotes the final value of the outer iteration counter when the augmented Lagrangian method of interest is aborted, then, by construction, **ALM** and **ALMr** have run precisely $300k$ iterations of the subproblem solver NADAM while **sALMr** has run precisely $10^4 + 300(k-1) \approx 300(k+32)$ iterations of it. Let us also emphasize that with an increasing outer iteration counter, the cost for one iteration of NADAM increases as well since the number of subboxes, taken into account for the computation of the stochastic gradient of the augmented Lagrangian function, is enlarged in each outer iteration; see Section 3.3. Thus, k is the decisive quantity for the running time of all three methods, and **sALMr** is, generally, slower than **ALM** and **ALMr** since the maximum number of outer iterations is set, for all three methods, to 30; see Section 3.2. For example, one run on the "Cameraman" image (see Figure 3.1) on an AMD Ryzen 5 Pro 5650U in our Python implementation took $64, 69$, and 81 minutes for **ALM**, **ALMr**, and **sALMr** with $k = 25$, $k = 27$, and $k = 30$, respectively.

3.5. Numerical results. In this section, we comment on the numerical behavior of the augmented Lagrangian schemes **ALM**, **ALMr**, and **sALMr** for the denoising of the three standard test images "Brain", "Butterfly", and "Cameraman", where the hyperparameters are chosen as described in the previous sections. The benchmark problems are created by applying Poisson noise to each of the three ground truth pictures $\hat{u}$. Thus, every pixel Z_i of the (discrete) noisy observation Z is randomly drawn according to

$$\forall i \in \{1, \dots, n\}^2 : \quad Z_i \sim \mathrm{Poi}(\hat{u}_i).$$

To obtain a reasonable benchmark collection of noisy test images, we created 10 noisy versions of each of the three test images mentioned above. The three test images of interest as well as some exemplary noisy versions of them can be found in Figure 3.1.

We pick r such that (1.8) and, consequently, (1.9) hold true with $\alpha = 0.1$. Typical examples of outputs associated with **ALM** can be found in Figure 3.2. The reconstructions show that the method yields reasonable results as convergence to a meaningful solution is observed. As mentioned before, using $\alpha = 0.1$ ensures the statistical guarantee (1.9) and thus leads by construction to a method tending to oversmoothing. This is clearly visible in Figure 3.2 but must be seen as a feature of the variational Poisson denoising method under

ETNA
Kent State University and
Johann Radon Institute (RICAM)

FIG. 3.1. *The three test images (top row) and noisy versions of the test images (bottom row). ©Copyrighted content. Creative commons license does not apply.*

FIG. 3.2. *Denoised versions of the test images obtained via **ALM**. ©Copyrighted content. Creative commons license does not apply.*

consideration. Let us note that the outputs of **ALMr** and **sALMr** are of comparable quality so we do not present the associated denoised images here.

To prevent the observed oversmoothing, one could apply the variational Poisson denoising approach with a smaller function r (shifted by a constant). For example, this can be seen in the denoised version of the "Cameraman" image in Figure 3.3, which, again, has been obtained with the aid of **ALM** but with decreased r. This modification implies that the statistical guarantee (1.9) is lost, but therefore the constraints are more restrictive and prevent oversmoothing. Note that a similar observation was made in [28]. It can be seen from Figure 3.3 that the corresponding reconstruction is less smooth but seems visually superior over the one in Figure 3.2. For our numerical comparison of the three methods **ALM**, **ALMr**, and **sALMr**, however, we stick to our original choice of r and α in order to preserve the idea behind variational Poisson denoising.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

FIG. 3.3. *Reconstruction of the "Cameraman" image with decreased r to reduce oversmoothing obtained via* ***ALM****. ©Copyrighted content. Creative commons license does not apply.*

Let us now start to document the comparison of the three augmented Lagrangian methods **ALM**, **ALMr**, and **sALMr** based on the performance measures mentioned in Section 3.4. Let us first inspect the performance profiles in Figure 3.4, which document the obtained objective function values, the total number of outer augmented Lagrangian iterations, as well as the value of the error measure and the size of the penalty parameter when the respective method is aborted. We observe that **ALM** computes the best objective function values for approximately

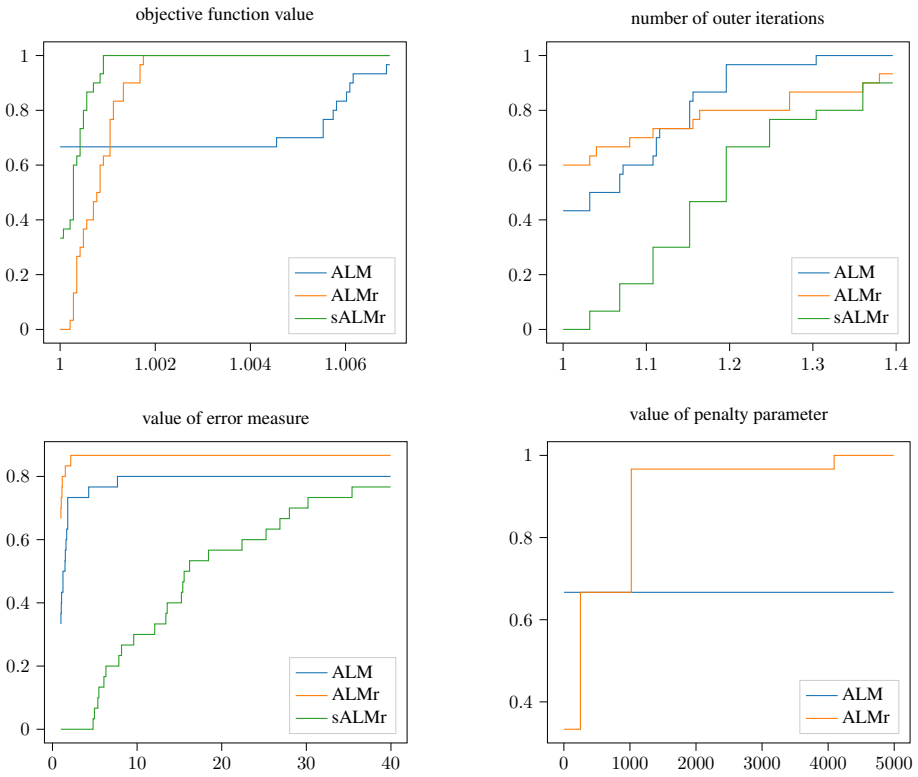

FIG. 3.4. *Performance profiles for the objective function value, the number of outer augmented Lagrangian iterations, the value of the error measure, and the value of the penalty parameter (from top left to bottom right).*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

67% of the test instances, but the scale of the t-axis underlines that **ALMr** and **sALMr** do not significantly fall behind **ALM** in this regard. Interestingly, **ALMr** and **sALMr** produce a point whose objective value is just 1‰ and 2‰ worse than the best computed function value for each test instance, respectively—a quality which **ALM** does not achieve. Regarding the total number of outer augmented Lagrangian iterations, we observe that **ALM** and **ALMr** perform best in about 40% and 60% of all test runs, respectively, while **sALMr** cannot match the other two methods in this regard. The performance profile also indicates that in about 90% of all test runs, the total number of outer iterations for any of the algorithms under consideration is at most 1.5 times as high as the smallest number of exploited outer iterations. Regarding the value of the error measure, there is a significant difference between **ALM** as well as **ALMr**, which perform similarly well with slight advantages for **ALMr**, and **sALMr** is outperformed by the other two methods. For example, the situation where the error measure is allowed to be 20 times higher than the best achieved value of the error measure is only covered in about 60% of all test runs for **sALMr**. The final value of the penalty parameter for **ALM** is smaller than the one for **ALMr** for about 68% of all test instances. However, while for its termination, **ALMr** reliably needs a penalty parameter which is at most $\gamma^6 = 4096$ times larger than the smallest necessary penalty parameter, i.e., at most 6 more enlargements of the penalty parameter in Step 6 of Algorithm 1 are necessary, a similar bound for **ALM** cannot be distilled from the associated performance profile.

Second, let us inspect the feasibility properties of the outputs of all these methods. In Figure 3.5, the percentage of violated constraints (with respect to the model (VPD)) is documented. Here, we observe that the smallest percentage of violated constraints is achieved

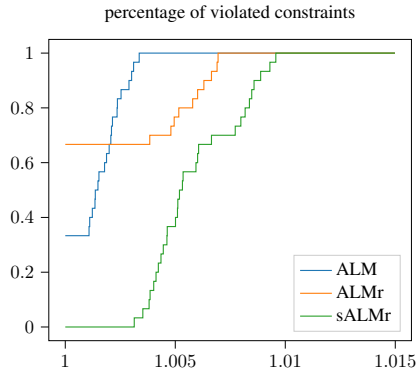

FIG. 3.5. *Performance profile for the percentage of violated constraints.*

by **ALMr** in about 68% of all test runs. However, **ALM** falls behind the smallest percentage of violated constraints only by a factor 1.003 in all test runs, and, anyhow, is in the lead for the other 32% of the test instances. Again, **sALMr** clearly cannot compete with **ALM** and **ALMr**. Still, it falls short the smallest percentage of violated constraints only by a factor 1.01, which is still an acceptable deviation. Note, however, that the percentage of violated constraints does not document how bad the violated constraints are missed. Therefore, let us inspect Figure 3.6 which reports the maximum and average relative (affine) constraint violation. Here, we observe that regarding the maximum and average relative constraint violation (with respect to the constraints of (VPD)), all methods perform in a good way with obvious advantages for **ALM** and **ALMr** in the settings of the average and maximum constraint violation, respectively. This is good news, as all three methods are, at their core, constructed to solve the model

ETNA
Kent State University and
Johann Radon Institute (RICAM)

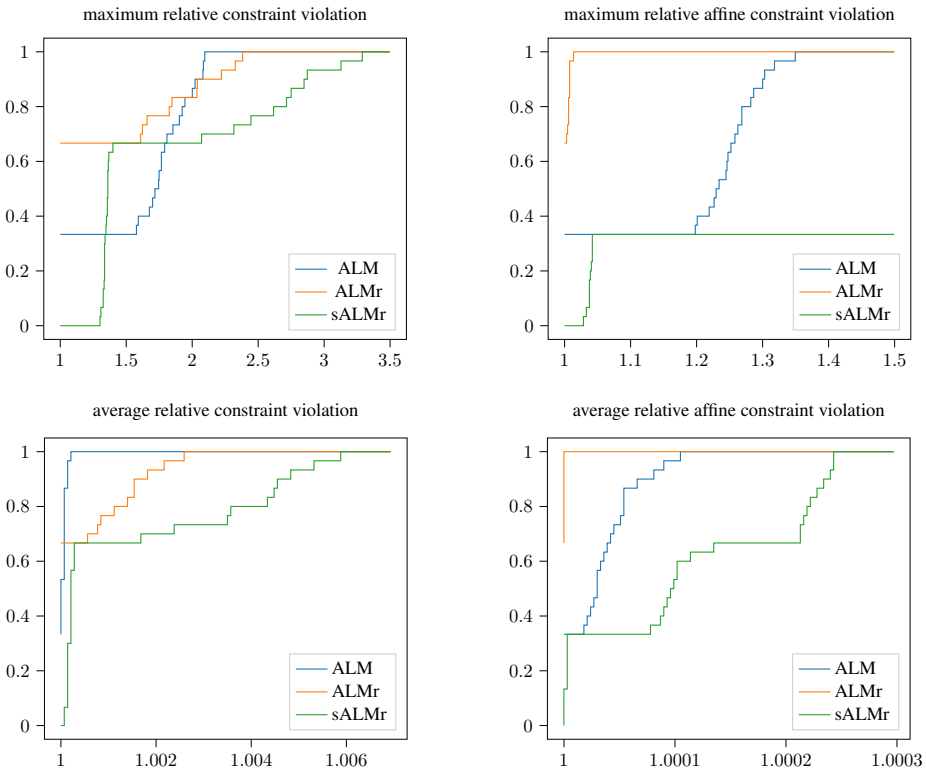

FIG. 3.6. *Performance profiles for the maximum relative constraint violation, the maximum relative affine constraint violation, the average relative constraint violation, and the average relative affine constraint violation (from top left to bottom right).*

problem (VPD), which comes along with the handling of the associated huge number of constraints. Let us now inspect the performance profiles for the maximum and average relative affine constraint violation (with respect to the constraints of (VPD_{aff})). One would expect that **ALMr** as well as **sALMr** outperform **ALM** in this regard as the former two algorithms directly work with the model formulation (VPD_{aff}) while the latter algorithm does not. This intuition is supported by our results only in parts. Indeed, we observe that **ALM** and **sALMr** both cannot challenge **ALMr** and that **ALM** performs significantly better than **sALMr**.

Let us present an interim summary of our numerical experiments. While **ALM** and **ALMr** both perform well on the set of benchmark problems, **sALMr** falls short both of these methods regarding the number of outer iterations, the final value of the error measure, and the feasibility aspects of the produced output. Furthermore, it has to be mentioned that, due to the fact that we need to run 10000 iterations of NADAM in order to solve the first augmented Lagrangian subproblem up to a reasonable precision to observe any convergence of **sALMr** later on (see Section 3.3), the latter is significantly slower than **ALM** and **ALMr**. We thus cannot attest **sALMr** a satisfying performance.

Next, we present some averaged numbers to analyze the behavior of the three methods on each of the three test images individually. Let us start with a documentation of the results associated with the test image "Brain" which can be found in Table 3.1. The performance profiles already indicated that all three methods compute points with related objective function value, and this is underlined by Table 3.1. Inspecting the total number of outer iterations

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 3.1
Averaged numbers for the test image "Brain".

performance measure	**ALM**	**ALMr**	**sALMr**
objective function value	$2.4729 \cdot 10^{8}$	$2.4592 \cdot 10^{8}$	$2.4583 \cdot 10^{8}$
number of outer iterations	29.3	26.5	30.0
value of error measure	$1.4998 \cdot 10^{1}$	$7.6307 \cdot 10^{-3}$	$3.8705 \cdot 10^{-2}$
value of penalty parameter	$3.4903 \cdot 10^{16}$	$1.0486 \cdot 10^{7}$	$4.0000 \cdot 10^{5}$
percentage of violated constraints	$9.4996 \cdot 10^{-5}$	$5.8363 \cdot 10^{-3}$	$8.4727 \cdot 10^{-3}$
max. rel. constraint violation	$3.7626 \cdot 10^{-3}$	$9.6763 \cdot 10^{-1}$	1.7180
max. rel. affine constraint violation	$2.2717 \cdot 10^{-5}$	$7.6307 \cdot 10^{-3}$	$3.8705 \cdot 10^{-2}$
av. rel. constraint violation	$1.1616 \cdot 10^{-8}$	$1.4712 \cdot 10^{-3}$	$4.0849 \cdot 10^{-3}$
av. rel. aff. constraint violation	$2.2768 \cdot 10^{-11}$	$6.7136 \cdot 10^{-7}$	$3.3754 \cdot 10^{-6}$

reveals an advantage of **ALMr** against **ALM** as well as **sALMr**, and similar observations can be made for the final value of the error measure and the penalty parameter. Let us note that **sALMr** exploits the predefined maximum total number of outer iterations in each run. It thus might be possible to achieve better results with this approach when allowing for further iterations. This, however, costs additional time, and we already mentioned above that **sALMr** is the slowest algorithm we are considering here. Regarding the feasibility aspects, **ALM** is clearly better than the other two methods. It has to be mentioned that there is a strong relation between these assessments. For the test image "Brain", which possesses broad areas of uniform color, we observe that, on many subboxes, the mean of the noisy image as well as the mean of the iterates are (very close to) zero. Numerically, this drives the value of the associated constraints in (VPD) and thus the error measure to ∞. In order to avoid this, Algorithm 1 enlarges the penalty parameter more often in this setting, compared to the one where (VPD_{aff}) is tackled, and this phenomenon cannot happen. At the end of the day, **ALM** then produces images which are much closer to feasibility.

Let us now inspect the averaged numbers for the results associated with the test images "Butterfly" and "Cameraman" in Tables 3.2 and 3.3, respectively.

TABLE 3.2
Averaged numbers for the test image "Butterfly".

performance measure	**ALM**	**ALMr**	**sALMr**
objective function value	$9.8967 \cdot 10^{8}$	$9.9047 \cdot 10^{8}$	$9.9007 \cdot 10^{8}$
number of outer iterations	27.9	26.4	30.0
value of error measure	$8.8339 \cdot 10^{-3}$	$7.3673 \cdot 10^{-3}$	$8.7777 \cdot 10^{-2}$
value of penalty parameter	$1.2452 \cdot 10^{5}$	$6.2076 \cdot 10^{7}$	$4.0000 \cdot 10^{5}$
percentage of violated constraints	$2.2192 \cdot 10^{-3}$	$7.4833 \cdot 10^{-5}$	$5.4026 \cdot 10^{-3}$
max. rel. constraint violation	$9.6253 \cdot 10^{-1}$	$1.8556 \cdot 10^{-2}$	$3.8648 \cdot 10^{-1}$
max. rel. affine constraint violation	$2.7286 \cdot 10^{-1}$	$7.3673 \cdot 10^{-3}$	1.4300
av. rel. constraint violation	$9.5834 \cdot 10^{-5}$	$7.6614 \cdot 10^{-8}$	$2.5851 \cdot 10^{-4}$
av. rel. aff. constraint violation	$4.4251 \cdot 10^{-5}$	$4.8633 \cdot 10^{-8}$	$2.2870 \cdot 10^{-4}$

Again, all three methods do not differ regarding the objective function value of the produced output. Inspecting the error measure reveals no big differences between **ALM** and **ALMr**, while **ALM** now terminates with a significantly smaller penalty parameter than **ALMr** does. In contrast, the feasibility measures look far better for **ALMr** than for **ALM**.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 3.3
Averaged numbers for the test image "Cameraman".

performance measure	**ALM**	**ALMr**	**sALMr**
objective function value	$1.1606 \cdot 10^{9}$	$1.1622 \cdot 10^{9}$	$1.1613 \cdot 10^{9}$
number of outer iterations	23.8	28.3	30.0
value of error measure	$6.8892 \cdot 10^{-3}$	$4.9436 \cdot 10^{-3}$	$1.2357 \cdot 10^{-1}$
value of penalty parameter	$3.0147 \cdot 10^{5}$	$1.4764 \cdot 10^{8}$	$4.0000 \cdot 10^{5}$
percentage of violated constraints	$1.9836 \cdot 10^{-3}$	$3.4649 \cdot 10^{-5}$	$4.1128 \cdot 10^{-3}$
max. rel. constraint violation	$7.7851 \cdot 10^{-1}$	$1.2058 \cdot 10^{-2}$	$3.4789 \cdot 10^{-1}$
max. rel. affine constraint violation	$2.6348 \cdot 10^{-1}$	$4.9436 \cdot 10^{-3}$	$7.9586 \cdot 10^{-1}$
av. rel. constraint violation	$1.1465 \cdot 10^{-4}$	$5.1053 \cdot 10^{-8}$	$1.6282 \cdot 10^{-4}$
av. rel. aff. constraint violation	$5.1804 \cdot 10^{-5}$	$2.5188 \cdot 10^{-8}$	$1.0092 \cdot 10^{-4}$

Interestingly, for the test image "Cameraman", **ALM** terminates after a notably smaller number of outer iterations than **ALMr**. As above, **sALMr** exploits the predefined maximum total number of outer iterations in each run. Regarding almost all aspects, **sALMr** cannot match the other two methods.

Summing up these results, we observe that both **ALM** and **ALMr** are reasonable strategies for the implementation of the Poisson denoising approach for image recovery while **sALMr** is not. As depicted above, the individual performance of **ALM** and **ALMr** depends on the structure of the underlying picture, so a general ranking between **ALM** and **ALMr** is not possible. This is rather remarkable since **ALM** is based on the nonlinear nonsmooth model problem (VPD) while **ALMr** builds on the seemingly simpler convex quadratic problem (VPD_{aff}).

4. Concluding remarks. In this paper, we discussed the numerical solution of the so-called variational Poisson denoising problem, which is a statistically motivated model for image denoising that possesses a huge number of constraints, with the aid of augmented Lagrangian methods. To tackle the original model (VPD), whose constraints are modeled with the aid of the extended real-valued Kullback–Leibler divergence and are thus nonsmooth, we suggested a rather general safeguarded augmented Lagrangian framework for fully nonsmooth problems in Banach spaces with finitely many inequality constraints, equality constraints within a Hilbert space setting, and additional abstract constraints, where the inequality and equality constraints are augmented. An associated derivative-free global convergence theory has been developed which applies in situations where the appearing subproblems can be solved to approximate global minimality, and the latter is likely to be possible in convex situations. Our results generalize related findings in (partially) smooth settings; see, e.g., [37, Section 4] or [39, Theorem 6.15]. For our analysis, we only relied on minimal requirements regarding semicontinuity properties of all involved data functions as (generalized) differentiation played no role, and this makes our results broadly applicable.

We also visualized that it is possible to equivalently reformulate the constraints in (VPD) as affine inequalities, but this procedure comes at the price of doubling the total number of constraints. After discretization, this new model (VPD_{aff}) is a convex quadratic optimization problem with a huge number of constraints and can be tackled with the standard safeguarded augmented Lagrangian framework. Furthermore, as the minimizers of (VPD_{aff}) are stationary, it is possible to solve the latter with the classical non-safeguarded version of the augmented Lagrangian method where the penalty parameter undergoes no evolution.

These three algorithmic approaches have been implemented to denoise a benchmark collection of noisy images obtained from the standard test images "Brain", "Butterfly", and

ETNA
Kent State University and
Johann Radon Institute (RICAM)

"Cameraman" by adding some random Poisson noise. On the one hand, our results document that the safeguarded augmented Lagrangian method performs comparably well when applied to (VPD) or its reformulation (VPD_{aff}). On the other hand, the classical non-safeguarded augmented Lagrangian method with a constant penalty parameter falls clearly short of the results obtained by the other two methods and thus cannot be considered a reasonable approach to solve the Poisson denoising model.

REFERENCES

[1] L. AMBROSIO, N. FUSCO, AND D. PALLARA, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, Oxford, 2000.
[2] T. ASPELMEIER, A. EGNER, AND A. MUNK, *Modern statistical challenges in high-resolution fluorescence microscopy*, Annu. Rev. Stat. Appl., 2 (2015), pp. 163–202.
[3] A. BENFENATI AND V. RUGGIERO, *Inexact Bregman iteration with an application to Poisson data reconstruction*, Inverse Problems, 29 (2013), Art. 065016, 31 pages.
[4] F. BENVENUTO, A. L. CAMERA, C. THEYS, A. FERRARI, H. LANTÉRI, AND M. BERTERO, *The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise*, Inverse Problems, 24 (2008), Art. 035016, 20 pages.
[5] M. BERTERO, P. BOCCACCI, G. DESIDERÀ, AND G. VICIDOMINI, *Image deblurring with Poisson data: from cells to galaxies*, Inverse Problems, 25 (2009), Art. 123006, 26 pages.
[6] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
[7] E. G. BIRGIN AND J. M. MARTÍNEZ, *Practical Augmented Lagrangian Methods for Constrained Optimization*, SIAM, Philadelphia, 2014.
[8] S. BONETTINI, A. BENFENATI, AND V. RUGGIERO, *Primal-dual first order methods for total variation image restoration in presence of Poisson noise*, in 2014 IEEE International Conference on Image Processing (ICIP), IEEE Conference Proceedings, Los Alamitos, 2014, pp. 4156–4160.
[9] S. BONETTINI AND V. RUGGIERO, *An alternating extragradient method for total variation-based image restoration from Poisson data*, Inverse Problems, 27 (2011), Art. 095001, 26 pages.
[10] ———, *On the convergence of primal–dual hybrid gradient algorithms for total variation image restoration*, J. Math. Imaging Vision, 44 (2012), pp. 236–253.
[11] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.
[12] E. BÖRGENS, C. KANZOW, P. MEHLITZ, AND G. WACHSMUTH, *New constraint qualifications for optimization problems in Banach spaces based on asymptotic KKT conditions*, SIAM J. Optim., 30 (2020), pp. 2956–2982.
[13] E. BÖRGENS, C. KANZOW, AND D. STECK, *Local and global analysis of multiplier methods for constrained optimization in Banach spaces*, SIAM J. Control Optim., 57 (2019), pp. 3694–3722.
[14] A. CHAMBOLLE AND T. POCK, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.
[15] X. CHEN, L. GUO, Z. LU, AND J. J. YE, *An augmented Lagrangian method for non-Lipschitz nonconvex programming*, SIAM J. Numer. Anal., 55 (2017), pp. 168–193.
[16] R. M. CORLESS, G. H. GONNET, D. E. G. HARE, D. J. JEFFREY, AND D. E. KNUTH, *On the Lambert W function*, Adv. Comput. Math., 5 (1996), pp. 329–359.
[17] A. DE MARCHI, X. JIA, C. KANZOW, AND P. MEHLITZ, *Constrained composite optimization and augmented Lagrangian methods*, Math. Program., 201 (2023), pp. 863–896.
[18] A. DE MARCHI AND P. MEHLITZ, *Local properties and augmented Lagrangians in fully nonconvex composite optimization*, J. Nonsmooth Anal. Optim., 5 (2024), Art. 12235, 36 pages.
[19] M. DEL ALAMO, H. LI, AND A. MUNK, *Frame-constrained total variation regularization for white noise regression*, Ann. Stat., 49 (2021), pp. 1318–1346.
[20] M. DEL ALAMO, H. LI, A. MUNK, AND F. WERNER, *Variational multiscale nonparametric regression: algorithms and implementation*, Algorithms (Basel), 13 (2020), Art. 296, 24 pages.
[21] N. K. DHINGRA, S. Z. KHONG, AND M. R. JOVANOVIĆ, *The proximal augmented Lagrangian method for nonsmooth composite optimization*, IEEE Trans. Automat. Contr., 64 (2019), pp. 2861–2868.
[22] E. D. DOLAN AND J. J. MORÉ, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.
[23] T. DOZAT, *Incorporating Nesterov momentum into ADAM*, Tech. Rep., International Conference on Learning Representations 2016, 2016.
[24] Q. DUAN, M. XU, Y. LU, AND L. ZHANG, *A smoothing augmented Lagrangian method for nonconvex, nonsmooth constrained programs and its applications to bilevel problems*, J. Ind. Manag. Optim., 15

ETNA
Kent State University and
Johann Radon Institute (RICAM)

(2019), pp. 1241–1261.

[25] L. FAN, F. ZHANG, H. FAN, AND C. ZHANG, *Brief review of image denoising techniques*, Vis. Comput. Ind. Biomed. Art, 2 (2019), 12 pages.

[26] M. A. T. FIGUEIREDO AND J. M. BIOUCAS-DIAS, *Restoration of Poissonian images using alternating direction optimization*, IEEE Trans. Image Process., 19 (2010), pp. 3133–3145.

[27] K. FRICK, P. MARNITZ, AND A. MUNK, *Statistical multiresolution Dantzig estimation in imaging: fundamental concepts and algorithmic framework*, Electron. J. Stat., 6 (2012), pp. 231–268.

[28] ———, *Statistical multiresolution estimation for variational imaging: with an application in Poisson-biophotonics*, J. Math. Imaging Vision, 46 (2013), pp. 370–387.

[29] M. GRASMAIR, H. LI, AND A. MUNK, *Variational multiscale nonparametric regression: smooth functions*, Ann. Inst. Henri Poincaré Probab. Stat., 54 (2018), pp. 1058–1097.

[30] L. GUO AND Z. DENG, *A new augmented Lagrangian method for MPCCs—theoretical and numerical comparison with existing augmented Lagrangian methods*, Math. Oper. Res., 47 (2022), pp. 1229–1246.

[31] N. T. V. HANG AND M. E. SARABI, *Local convergence analysis of augmented Lagrangian methods for piecewise linear-quadratic composite optimization problems*, SIAM J. Optim., 31 (2021), pp. 2665–2694.

[32] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888.

[33] T. HOHAGE AND F. WERNER, *Inverse problems with Poisson data: statistical regularization theory, applications and algorithms*, Inverse Problems, 32 (2016), Art. 093001, 56 pages.

[34] K. ITO AND K. KUNISCH, *Lagrange Multiplier Approach to Variational Problems and Applications*, SIAM, Philadelphia, 2008.

[35] X. JIA, C. KANZOW, P. MEHLITZ, AND G. WACHSMUTH, *An augmented Lagrangian method for optimization problems with structured geometric constraints*, Math. Program., 199 (2023), pp. 1365–1415.

[36] C. KANZOW AND D. STECK, *An example comparing the standard and safeguarded augmented Lagrangian methods*, Oper. Res. Lett., 45 (2017), pp. 598–603.

[37] C. KANZOW, D. STECK, AND D. WACHSMUTH, *An augmented Lagrangian method for optimization problems in Banach spaces*, SIAM J. Control Optim., 56 (2018), pp. 272–291.

[38] C. KÖNIG, A. MUNK, AND F. WERNER, *Multidimensional multiscale scanning in exponential families: limit theory and statistical consequences*, Ann. Statist., 48 (2020), pp. 655–678.

[39] A. Y. KRUGER AND P. MEHLITZ, *Optimality conditions, approximate stationarity, and applications—a story beyond Lipschitzness*, ESAIM Control Optim. Calc. Var., 28 (2022), Art. 42, 35 pages.

[40] T. LE, R. CHARTRAND, AND T. J. ASAKI, *A variational approach to reconstructing images corrupted by Poisson noise*, J. Math. Imaging Vision, 27 (2007), pp. 257–263.

[41] J. LI, Z. SHEN, R. YIN, AND X. ZHANG, *A reweighted ℓ^2 method for image restoration with Poisson and mixed Poisson-Gaussian noise*, Inverse Probl. Imaging, 9 (2015), pp. 875–894.

[42] Z. LU, Z. SUN, AND Z. ZHOU, *Penalty and augmented Lagrangian methods for constrained DC programming*, Math. Oper. Res., 47 (2022), pp. 2260–2285.

[43] A. MUNK, K. PROKSCH, H. LI, AND F. WERNER, *Photonic imaging with statistical guarantees: from multiscale testing to multiscale estimation*, in Nanoscale Photonic Imaging, T. Salditt, A. Egner, and D. R. Luke, eds., Springer, Cham, 2020, pp. 283–312.

[44] R. T. ROCKAFELLAR, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555–562.

[45] ———, *Convergence of augmented Lagrangian methods in extensions beyond nonlinear programming*, Math. Program., 199 (2023), pp. 375–420.

[46] J. SHARPNACK AND E. ARIAS-CASTRO, *Exact asymptotics for the scan statistic and fast alternatives*, Electron. J. Stat., 10 (2016), pp. 2641–2684.

[47] D. L. SNYDER, C. W. HELSTROM, A. D. LANTERMAN, M. FAISAL, AND R. L. WHITE, *Compensation for readout noise in CCD images*, J. Opt. Soc. Am. A, 12 (1995), pp. 272–283.

[48] D. L. SNYDER, A. M. HAMMOUD, AND R. L. WHITE, *Image recovery from data acquired with a charge-coupled-device camera*, J. Opt. Soc. Am. A, 10 (1993), pp. 1014–1023.

[49] D. N. H. THANH, V. B. S. PRASATH, AND L. M. HIEU, *A review on CT and X-ray images denoising methods*, Informatica (Ljubl.), 43 (2019), pp. 151–159.

[50] A. B. TSYBAKOV, *Introduction to Nonparametric Estimation*, Springer, New York, 2009.

[51] Y. VARDI, L. A. SHEPP, AND L. KAUFMAN, *A statistical model for positron emission tomography*, J. Amer. Statist. Assoc., 80 (1985), pp. 8–37.

[52] M. XU, J. J. YE, AND L. ZHANG, *Smoothing augmented Lagrangian method for nonsmooth constrained optimization problems*, J. Global Optim., 62 (2015), pp. 675–694.

[53] C. ZĂLINESCU, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, 2002.

[54] R. ZANELLA, P. BOCCACCI, L. ZANNI, AND M. BERTERO, *Efficient gradient projection methods for edge-preserving removal of Poisson noise*, Inverse Problems, 25 (2009), Art. 045010, 24 pages.

[55] L. ZANNI, A. BENFENATI, M. BERTERO, AND V. RUGGIERO, *Numerical methods for parameter estimation in Poisson data inversion*, J. Math. Imaging Vision, 52 (2015), pp. 397–413.