PRIMAL–DUAL PROXIMAL SPLITTING AND GENERALIZED CONJUGATION IN NON-SMOOTH NON-CONVEX OPTIMIZATION

Christian Clason* Stanislav Mazurenko† Tuomo Valkonen‡

2020–02–12

Abstract We demonstrate that difficult non-convex non-smooth optimization problems, such as Nash equilibrium problems and anisotropic as well as isotropic Potts segmentation model, can be written in terms of generalized conjugates of convex functionals. These, in turn, can be formulated as saddle-point problems involving convex non-smooth functionals and a general smooth but non-bilinear coupling term. We then show through detailed convergence analysis that a conceptually straightforward extension of the primal–dual proximal splitting method of Chambolle and Pock is applicable to the solution of such problems. Under sufficient local strong convexity assumptions of the functionals – but still with a non-bilinear coupling term – we even demonstrate local linear convergence of the method. We illustrate these theoretical results numerically on the aforementioned example problems.

1 INTRODUCTION

This work is concerned with the numerical solution of non-smooth non-convex saddle-point problems of the form

$$\min_{x\in X}\max_{y\in Y} G(x) + K(x,y) - F^*(y), \tag{1.1}$$

where $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$ are (possibly non-smooth) proper, convex and lower semicontinuous functionals on Hilbert spaces X and Y, and $K : X \times Y \to \mathbb{R}$ is smooth but may be non-convex-concave. Such problems arise in many areas of optimal control, inverse problems, and imaging; we will treat two specific examples below. To find a critical point for (1.1), we propose the *generalized primal–dual proximal splitting* (GPDPS) method:

*Faculty of Mathematics, University Duisburg-Essen, 45117 Essen, Germany (christian.clason@uni-due.de, ORCID: 0000-0002-9948-8426)

†Loschmidt Laboratories, Masaryk University, Brno, Czechia; *previously* Department of Mathematical Sciences, University of Liverpool, United Kingdom (stan.mazurenko@gmail.com, ORCID: 0000-0003-3659-4819)

‡ModeMat, Escuela Politécnica Nacional, Quito, Ecuador *and* Department of Mathematics and Statistics, University of Helsinki, Finland; *previously* Department of Mathematical Sciences, University of Liverpool, United Kingdom (tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572)

Algorithm 1.1 (GPDPS). Given an initial iterate (x^0, y^0) and rules for steplengths $\tau_i, \omega_i, \sigma_i > 0$, iterate:

$$\begin{aligned} x^{i+1} &:= \operatorname{prox}_{\tau_i G}(x^i - \tau_i K_x(x^i, y^i)), \\ \bar{x}^{i+1} &:= x^{i+1} + \omega_i(x^{i+1} - x^i), \\ y^{i+1} &:= \operatorname{prox}_{\sigma_{i+1} F^*}(y^i + \sigma_{i+1} K_y(\bar{x}^{i+1}, y^i)), \end{aligned}$$

where $\operatorname{prox}_{\tau_i G}(v) = (I + \tau_i \partial G)^{-1}(v)$ is the proximal mapping for G; and K_x, K_y are the partial Fréchet derivatives of K with respect to x and y. A main result of this work is that under suitable conditions on the steplength parameters τ_i, σ_i, and ω_i, this algorithm converges weakly to a critical point of (1.1); see Theorem 6.1. Furthermore, if ∂G and/or ∂F^* is strongly metrically subregular at the saddle point (in particular, if G and/or F^* are strongly convex), we show optimal convergence rates for the standard acceleration strategies; see Theorems 6.3 and 6.4.

In addition, we demonstrate in this work how through a suitable reformulation this method can be applied to the following two non-trivial applications:

(i) *elliptic Nash equilibrium problems*, where $K(x, y)$ is the so-called *Nikaido–Isoda* function encoding the Nash equilibrium [25, 29, 38]; see Section 2.1 for details.

(ii) *(Huber-regularized) ℓ^0-TV denoising* (also referred to as the *Potts model*) [18, 33, 34], where $K(x, y)$ is used to express the non-convex Potts functional as the *generalized K-conjugate* of a convex indicator function; see Section 2.2 for details.

In particular, the second example demonstrates how the proposed method can be used to solve (some) non-convex non-smooth problems by reformulating in them in terms of a convex but non-smooth functional and a smooth but non-convex coupling term. (We stress, however, that we do not claim that this approach is superior to state-of-the-art problem-specific approaches such as the ones mentioned in the cited works for the specific problems; such an investigation is left for the future.)

Related literature. Our approach is obviously motivated by the well-known primal–dual proximal splitting (PDPS) method of Chambolle and Pock [8] for convex optimization problems of the form $\min_{x \in X} F(Ax) + G(x)$ for $F : Y \to \overline{\mathbb{R}}$ proper, convex, and lower semicontinuous and $A : X \to Y$ linear. The method is based on the equivalent reformulation as the saddle-point problem

$$\min_{x \in X} \max_{y \in Y} G(x) + \langle Ax, y \rangle - F^*(y) \tag{1.2}$$

where F^* is the Fenchel conjugate of F. Several other alternative techniques for such optimization problems have also been developed, e.g., using smoothing schemes [28] or a proximal alternating predictor corrector [13]. This approach was extended to allow for nonlinear but Fréchet differentiable A in [35]. Later work [10, 12] applied this to non-convex PDE-constrained optimization problems and derived accelerated variants.

In a broader context, generalized convex conjugation has been studied for many decades with applications in economics, see, e.g., [15, 26, 32] and the references therein. Algorithms for the solution of general saddle-point problems $\min_x \max_y f(x,y)$ have been considered in several seminal papers. In particular, a prox-type method was suggested in [27] for $C^{1,1}$ convex–concave functions yielding a $O(1/N)$ rate of convergence for an ergodic version of the gap $\max_{y'\in Y} f(x,y') - \min_{x'\in X} f(x',y)$. These results were further extended to allow non-smooth functions in the Mirror Descent method [22], demonstrating a $O(1/\sqrt{N})$ rate of convergence for the ergodic gap although with a vanishing step size for large N. The authors also considered an acceleration of the Mirror Proximal method for the case when the gradient map of f can be split into a Lipschitz-continuous part and a monotone operator [23]. The latter was assumed "simple" in the sense that a solution to a specific variational inequality could be found relatively efficiently. As a result, the authors obtained an $O(1/N)$ rate of convergence with a possibility for improvement to $O(1/N^2)$ for a strongly concave f. Finally, the reformulation of (1.1) with a bilinear K as a monotone inclusion problem was considered in [21].

Algorithms applicable to (1.1) with a genuinely nonlinear K have only started to appear in literature relatively recently. An abstract convergence result was obtained for an inexact regularized Gauss–Seidel method in [3]. In [20], the authors considered saddle-point representable functions and arrived at a very similar structure to (1.1); specifically, they reformulated this problem as a smooth linearly-constrained saddle point problem by moving the non-smooth terms into the problem domain and applied the Mirror Proximal algorithm mentioned earlier, with a smooth cost function and the $O(1/N)$ convergence rate [27]. Following [21], Kolossoski and Monteiro [24] developed a non-Euclidean hybrid proximal extragradient for G and F^* Bregman distances, and K general convex–concave.

The case of a general convex–concave K in (1.1) (which therefore becomes an overall convex–concave problem) has been recently studied in [19]. Finally, problems for general sufficiently smooth $K(x,y)$ were considered in [5] in conjunction with a variant of ADMM; however, no proofs of convergence were given in the general case.

Organization. To motivate our approach, we start with a more detailed description of the above-mentioned example problems and their reformulation as a saddle-point problem of the form (1.1) in the next Section 2. (This section can be skipped by readers only interested in the convergence analysis for the general Algorithm 1.1.) The following Section 3 then collects basic notation and definitions as well as the fundamental assumptions that will be used throughout the following. We then study the convergence and convergence rates of Algorithm 1.1 in Sections 4 to 6. More precisely, in Section 4 we derive a basic convergence estimate using the "testing" framework introduced in [36, 37] for the study of preconditioned proximal point methods. The results and assumptions depend on the iterates staying in a local neighborhood of a solution. In Section 5 we therefore derive conditions on the steplength parameters and initial iterate that ensure that the iterates do not escape from a local neighborhood. Afterwards, we provide in Section 6 exact steplength rules for Algorithm 1.1 together with respective weak convergence or convergence rate results: linear under sufficient strong convexity of G and F^*, and "accelerated" $O(1/N)$ or $O(1/N^2)$ rates with somewhat lesser assumptions. Finally, we illustrate the applicability and performance of the proposed approach applied to our two example problems in Section 7.

Appendices A to C contain further technical results on the assumptions required for convergence, in particular verifying them for the Huber-regularized ℓ^0-TV denoising example.

2 APPLICATIONS

Before we begin our analysis of the convergence of Algorithm 1.1, we motivate its generality by discussing two examples of practically relevant problems that can be cast in the form (1.1) and which will be used to numerically illustrate the behavior of the algorithm in Section 7. The idea in each case is to write a *non-convex* functional F as the generalized K-conjugate of a *convex* functional F^*, i.e.,

$$F(x) = \sup_{y \in Y} K(x, y) - F^*(y)$$

for a suitable K (depending on F).

2.1 ELLIPTIC NASH EQUILIBRIUM PROBLEMS

Our first example is the reformulation of Nash equilibrium problems using the Nikaido–Isoda function following [38]. Consider a non-cooperative game of $n \in \mathbb{N}$ players, each of which has a strategy $x_k \in X_k \subset \mathbb{R}$ and a payout function $\phi_k : \mathbb{R}^n \to \mathbb{R}$. For convenience, we introduce the vector $x \in \mathbb{R}^n$ of strategies and the notation

$$(x_{-k}|z) := (x_1, \dots, x_{k-1}, z, x_{k+1}, \dots x_n) \qquad (1 \le k \le n,\ z \in \mathbb{R})$$

for the vector where player k changes their strategy x_k to z. We also set $X := X_1 \times \cdots \times X_n$. A vector $x^* \in X$ of strategies is then a Nash equilibrium if

$$\phi_k(x^*) = \phi_k(x^*_{-k}|x^*_k) = \min_{z \in \mathbb{R}} \phi_k(x^*_{-k}|z) \qquad (1 \le k \le n). \tag{2.1}$$

We now introduce the Nikaido–Isoda function [29] (also called the Ky Fan function [17])

$$\Psi(x, y) = \sum_{k=1}^{n} (\phi_k(x_{-k}|x_k) - \phi_k(x_{-k}|y_k)) \qquad (x, y \in X)$$

as well as the optimum response function

$$V(x) = \max_{y \in X} \Psi(x, y) \qquad (x \in X). \tag{2.2}$$

It follows from [38, Thm. 2.2] that $x^* \in X$ is a Nash equilibrium if and only if it is a minimizer of V. Using the indicator function of the set $X \subset \mathbb{R}^n$ defined by

$$\delta_X(x) = \begin{cases} 0 & \text{if } x \in X, \\ \infty & \text{if } x \notin X, \end{cases}$$

we see that the generally non-convex response function V is the Ψ-preconjugate of the convex functional δ_X and can characterize a Nash equilibrium $x^* \in X$ as the solution to the saddle-point problem

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} \delta_X(x) + \Psi(x, y) - \delta_X(y).$$

We can therefore solve the Nash equilibrium problem (2.1) by applying Algorithm 1.1 to

$$K(x, y) = \Psi(x, y), \qquad F^* = G = \delta_X.$$

In Section 7.1, we illustrate this exemplarily for the two-player elliptic Nash equilibrium problem from [6].

Remark 2.1. *If the set X_k of feasible strategies for each player depends on the strategies of the other players (i.e., $X_k = X_k(x_{-k})$), (2.1) becomes a* generalized Nash equilibrium problem (GNEP)*; see the survey [16] and the literature cited therein. If for all k*

$$X_k(x_{-k}) = \{x_k \in \mathbb{R}^n : (x_{-k}|x_k) \in Z\} \qquad (1 \le k \le n)$$

for some closed and convex set $Z \subset \mathbb{R}^n$, the GNEP is called jointly convex*. In this case, minimization of (2.2) is no longer an equivalent characterization but defines a* variational equilibria *[31]; every variational equilibrium is a generalized Nash equilibrium but not vice versa, see, e.g., [16, Thm. 3.9]. Hence Algorithm 1.1 can also be applied to compute (some if not all) solutions to jointly convex GNEPs.*

2.2 HUBER–POTTS DENOISING

Our next example is concerned with (Huber-regularized) ℓ^0-TV denoising or segmentation, also referred to as *Potts model*. Let $f \in \mathbb{R}^{N_1 \times N_2}$, $N_1, N_2 \in \mathbb{N}$, be a given noisy or to be segmented image. We then search for the denoised or segmented image as the solution to

$$\min_{x \in \mathbb{R}^{N_1 \times N_2}} \frac{1}{2\alpha}\|x - f\|^2 + \|D_h x\|_{p,0}, \tag{2.3}$$

for a regularization parameter $\alpha \ge 0$ (which we write in front of the discrepancy term to simplify the computations), the discrete gradient $D_h : \mathbb{R}^{N_1 \times N_2} \to \mathbb{R}^{N_1 \times N_2 \times 2}$, and the vectorial ℓ^0-seminorm

$$\|z\|_{p,0} := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left|\left(|z_{ij1}|_0, |z_{ij2}|_0\right)\right|_p, \qquad \text{where } |t|_0 = \begin{cases} 0 & \text{if } t = 0, \\ 1 & \text{if } t \neq 0, \end{cases} \tag{2.4}$$

and $|\cdot|_p$ for $p \in [1, \infty]$ is the usual p-norm on $\mathbb{R}^2$; we will discuss the choice of p in detail below. Clearly, $\|\cdot\|_{p,0}$ is a non-convex functional for any $p \in [1, \infty]$. Let us briefly comment on the use of ℓ^0-TV as a regularizer in imaging. Intuitively, the functional in (2.4) applied to the discrete gradient counts the number of jumps of the image value between neighboring pixels; it can therefore be expected that minimizers are piecewise constant, and that jumps are penalized even more strongly than by the (convex) total variation model.

To motivate our approach, we first consider a simple scalar (lower semicontinuous) step function, i.e., we consider for $(0, \infty) \subset \mathbb{R}$ the corresponding *characteristic function*

$$\chi_{(0,\infty)}(t) = \begin{cases} 0 & \text{if } t \le 0, \\ 1 & \text{if } t > 0. \end{cases} \tag{2.5}$$

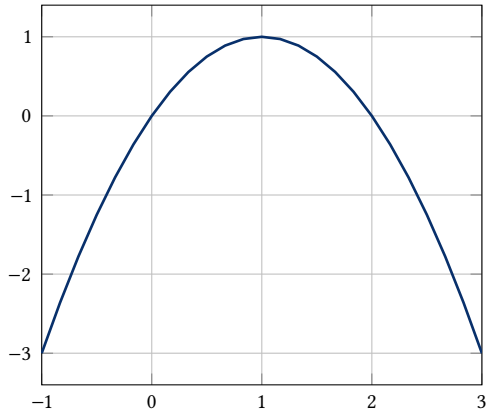


Figure 1: plot of ρ from (2.7)

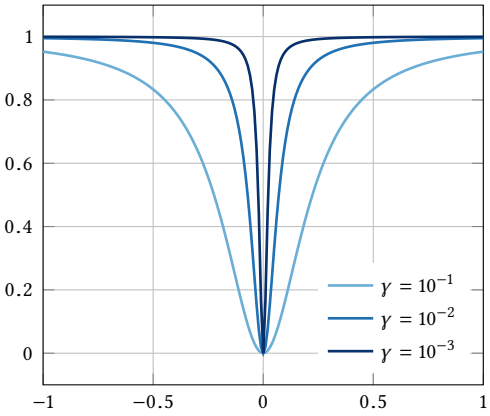


Figure 2: plot of $|t|_\gamma$ for different values of γ

To write this non-convex function as the generalized preconjugate of a convex function, let $\rho : \mathbb{R} \to \mathbb{R}$ satisfy $\rho(0) = 0$, $\sup_{t \le 0} \rho(t) = 0$, and $\sup_{t>0} \rho(t) = 1$. Then a simple case distinction shows that

$$\chi_{(0,\infty)}(t) = \sup_{s \ge 0} \rho(st) = \sup_{s \in \mathbb{R}} \rho(st) - \delta_{[0,\infty)}(s). \tag{2.6}$$

Setting $\kappa(s,t) := \rho(st)$, we thus obtain that $\chi_{(0,\infty)}$ is the κ-preconjugate of the convex indicator function $\delta_{[0,\infty)}$. One possible choice for ρ is $\rho = \chi_{(0,\infty)}$; however, we require ρ to be smooth in order to apply Algorithm 1.1. A better choice is therefore

$$\rho(t) = 2t - t^2, \qquad (t \in \mathbb{R}), \tag{2.7}$$

see Figure 1, which has the advantage that the supremum in (2.6) is always attained at a finite $s \ge 0$. We will use this choice from now on.

Noting that $|t|_0 = \chi_{\{0\}}(t)$, we can proceed similarly by case distinction to write

$$|t|_0 = \sup_{s \in \mathbb{R}} \rho(st) = \sup_{s \in \mathbb{R}} \rho(st) - 0,$$

i.e., for $\kappa(s,t) = \rho(st)$ as above, $|\cdot|_0$ is the κ-preconjugate of the zero function $f^* \equiv 0$. In practice, it may be useful to add Huber regularization, i.e., replace f^* by $f^*_\gamma := f^* + \frac{\gamma}{2}|\cdot|^2 = \frac{\gamma}{2}|\cdot|^2$ for some $\gamma > 0$. Using the fact that f^*_γ and our choice (2.7) are differentiable, an elementary calculus argument shows that the corresponding preconjugate is

$$|t|_\gamma := \sup_{s \in \mathbb{R}} \rho(st) - \frac{\gamma}{2}|s|^2 = \frac{2t^2}{2t^2 + \gamma},$$

which is a still non-convex approximation of $|t|_0$, see Figure 2.

We now turn to the vectorial ℓ^0 seminorm, where we distinguish between $p \in [1, \infty]$.

The case $p = 1$. With this choice, (2.4) reduces to

$$\|z\|_{1,0} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{2} |z_{ijk}|_0,$$

which is the most common choice for the Potts model found in the literature. Here, the Potts functional $\|D_h x\|_{1,0}$ counts for each pixel (i, j) the jumps across each edge of the pixel separately, i.e., the contribution of each pixel is either 0 (no jump), 1 (jump in either horizontal or vertical direction), or 2 (jump in both directions). We thus refer (in a slight abuse of terminology) to this case as the *anisotropic Potts model*.

Since this functional is completely separable, we can apply the above scalar approach componentwise by taking

$$\kappa_1(z, y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{2} \rho(z_{ijk} y_{ijk}) \tag{2.8}$$

such that $F = \| \cdot \|_{1,0}$ is the κ_1-preconjugate of the zero function $F^* \equiv 0$. Correspondingly, the Huber regularization of F is given by

$$F_\gamma(z) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{k=1}^{2} |z_{ijk}|_\gamma .$$

The case $p = \infty$. Now (2.4) reduces to

$$\|z\|_{\infty,0} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \max\left\{|z_{ij1}|_0, |z_{ij2}|_0\right\} .$$

Here, each pixel contributes to the Potts functional only once, even if there is a jump across both edges. Since a simple case distinction shows that $\max\{|a|_0, |b|_0\} = ||(a, b)|_p|_0$ for any $a, b \in \mathbb{R}$ and $p \in [1, \infty]$, this case is equivalent to

$$|||z|||_{0,p} := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left| |(z_{ij1}, z_{ij2})|_p \right|_0$$

for any $p \in [1, \infty]$, which leads to an alternate definition of the Potts functional sometimes found in the literature. We refer to this case as the *isotropic Potts model*.

This functional is only separable with respect to the pixel coordinates (i, j) but not with respect to k. We thus extend our preconjugation approach to $\mathbb{R}^2$ by observing for $t \in \mathbb{R}^2$ that

$$||t|_2|_0 = \sup_{s \in \mathbb{R}} \rho(\langle s, t \rangle) = \sup_{s \in \mathbb{R}} \rho(s_1 t_1 + s_2 t_2)$$

since for $t = 0$, $\rho(\langle s, t \rangle) = 0$ for all $s \in \mathbb{R}^2$, while for $t_1 \neq 0$ or $t_2 \neq 0$, the supremum will be attained at 1 by the choice of ρ. Setting

$$\kappa_\infty(z, y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \rho\left(z_{ij1} y_{ij1} + z_{ij2} y_{ij2}\right) \tag{2.9}$$

makes $F = \|\cdot\|_{\infty,0}$ again the κ_∞-preconjugate of the zero function $F^* \equiv 0$. The corresponding Huber regularization can be once more computed by elementary calculus as

$$F_\gamma(z) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left| |(z_{ij1}, z_{ij2})|_2 \right|_\gamma .$$

The case $p \in (1, \infty)$. In principle, one could proceed as for $p = \infty$ by constructing a function $\rho_p : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ with

$$\sup_{s \in \mathbb{R}^2} \rho_p(s,t) = \begin{cases} 0 & \text{if } t = 0, \\ 1 & \text{if } t \neq 0,\ t_1 t_2 = 0, \\ 2^{1/p} & \text{if } t \neq 0,\ t_1 t_2 \neq 0, \end{cases}$$

and setting $\kappa_p(s,t) = \rho_p(s,t)$. However, since the corresponding Potts functional only differs from the case $p = 1$ by the relative contribution of pixels with jumps in both directions and $2^{1/p} \to 1$ for $p \to \infty$, we will only consider the extremal cases $p = 1$ and $p = \infty$.

In all cases, we can apply Algorithm 1.1 to

$$K(x,y) = \kappa_p(D_h x, y), \qquad G(x) = \frac{1}{2\alpha}\|x - f\|^2, \qquad F_\gamma^*(y) = \frac{\gamma}{2}\|y\|^2$$

for $p \in [1, \infty]$ and $\gamma \geq 0$. We illustrate the application of Algorithm 1.1 for $p \in \{1, \infty\}$ and $\gamma > 0$ in Section 7.2.

Remark 2.2. *We can also apply this approach for $|t|^q$ with $q \in (0,1)$ using the same ρ as above, writing*

$$|t|^q = \sup_{s \in \mathbb{R}} \kappa(t,s) \quad \textit{for} \quad \kappa(t,s) := |t|^q \rho(st),$$

as $\rho(st) = 0$ if $t = 0$ and attains the maximal value 1 otherwise. However, $\kappa(t,s)$ is not C^2; we can achieve that by instead writing

$$|t|^q = \sup_{s \in \mathbb{R}} \kappa(t,s) \quad \textit{for} \quad \kappa(t,s) := |t|^q \rho(|st|^2).$$

3 NOTATION AND ASSUMPTIONS

We start the development of our proposed method by introducing the necessary notation and overall assumptions. Throughout the rest of this paper, we write $\mathbb{L}(X;Y)$ for the space of bounded linear operators between Hilbert spaces X and Y. In what follows, we let x and y denote elements of X and Y, respectively, and denote by u a pair $(x,y) \in X \times Y$. For brevity, we will also use this notation for similar tuples, e.g., $u^i := (x^i, y^i)$, without explicit introduction in each case.

For any Hilbert space, I is the identity operator, $\langle x, x' \rangle$ is the inner product in the corresponding space, and $\mathbb{B}(x,r)$ is the closed unit ball of the radius r at x. If $H : X \rightrightarrows X$ is a set-valued map, we will frequently use the concise notation

$$\langle H(x), \widetilde{x} \rangle := \{ \langle w, \widetilde{x} \rangle : w \in H(x) \}$$

as well as, e.g.,

$$0 \le \langle H(x), \widetilde{x} \rangle$$

if the corresponding relation holds for *all* $w \in H(x)$.

For self-adjoint $T, S \in \mathbb{L}(X;Y)$, the inequality $T \ge S$ means $T - S$ is positive semidefinite. If $T \in \mathbb{L}(X;X)$ is self-adjoint, we further set $\langle x, x' \rangle_T := \langle Tx, x' \rangle$, and $\|x\|_T := \sqrt{\langle x, x \rangle_T}$ (which define an inner product and a norm in X, respectively, if T is in addition positive definite). In this case, $T \ge S$ implies that $\|x\|_T \ge \|x\|_S$ for all $x \in X$.

We also recall that K_x and K_y denote the partial Fréchet derivatives of a continuosly differentiable operator K with respect to the given variable.

Throughout this paper, we make the following fundamental assumptions on (1.1).

Assumption 3.1. The functionals $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$ are convex, proper, and lower semicontinuous. Furthermore,

(i) there exist a constant $\gamma_G \in \mathbb{R}$ and a neighborhood $\mathcal{X}_G$ of $\widehat{x}$ such that

$$\langle \partial G(x) + K_x(\widehat{x}, \widehat{y}), x - \widehat{x} \rangle \ge \gamma_G \|x - \widehat{x}\|^2 \qquad (x \in \mathcal{X}_G); \tag{3.1}$$

(ii) there exist a constant $\gamma_{F^*} \in \mathbb{R}$ and a neighborhood $\mathcal{Y}_{F^*}$ of $\widehat{y}$ such that

$$\langle \partial F^*(y) - K_y(\widehat{x}, \widehat{y}), y - \widehat{y} \rangle \ge \gamma_{F^*} \|y - \widehat{y}\|^2 \qquad (y \in \mathcal{Y}_{F^*}). \tag{3.2}$$

Let us comment on this assumption. First, since the subgradients ∂G and ∂F^* of convex, proper, and lower semicontinuous functionals are maximally monotone operators [4, Theorem 20.25], Assumption 3.1 always holds with $\gamma_G = \gamma_{F^*} = 0$. This is already sufficient for showing weak convergence of Algorithm 1.1; see Theorem 6.1. For strong convergence with rates, however, we (as usual in nonlinear optimization) need a local superlinear growth condition near the solution that requires taking γ_G and/or γ_{F^*} strictly positive (unless we can compensate by better properties of K through Assumption 3.2 below); see Theorems 6.3 and 6.4. In this case, Assumption 3.1 (i), for example, coincides with *strong metric subregularity* of ∂G; see [1, 2]. This property holds (at *any* $\hat{x}$ and $\hat{w} \in \partial G(\hat{x})$) whenever G is strongly convex; however, it is a strictly weaker property since we only require it to hold at a *specific* $\hat{x}$ and $\hat{w} = -K_x(\widehat{x}, \widehat{y})$ arising from the first-order necessary optimality conditions (4.1) below. (For example, ∂g for $g(x) = |x|$ is strongly metrically subregular at $x = 0$ for $w \in (-1, 1)$ – but not at $w \in \{-1, 1\}$ – although g is not strongly convex.)

Assumption 3.2. The functional $K(x, y) \in C^1(X \times Y)$ and there exist $\rho_x, \rho_y > 0$ such that for all

$$u, u' \in \mathcal{U}(\rho_x, \rho_y) := (\mathbb{B}(\widehat{x}, \rho_x) \cap \mathcal{X}_G) \times (\mathbb{B}(\widehat{y}, \rho_y) \cap \mathcal{Y}_{F^*}), \tag{3.3}$$

the following properties hold:

(i) (second partial derivatives) The second partial derivatives $K_{xy}(u)$ and $K_{yx}(u)$ exist and satisfy $K_{xy}(u) = [K_{yx}(u)]^*$.

(ii) (locally Lipschitz gradients) For some functions $L_x(y), L_y(x) \geq 0$ and a constant $L_{yx} \geq 0$,

$$\|K_x(x', y) - K_x(x, y)\| \leq L_x(y)\|x' - x\|, \quad \|K_{yx}(x', y) - K_{yx}(x, y)\| \leq L_{yx}\|x' - x\|,$$
$$\|K_y(x, y') - K_y(x, y)\| \leq L_y(x)\|y' - y\|.$$

(iii) (locally bounded gradient) There exists $R_K > 0$ with $\sup_{u \in \mathcal{U}(\rho_x, \rho_y)} \|K_{xy}(x, y)\| \leq R_K$.

(iv) (three-point condition) There exist $\theta_x, \theta_y > 0$, $\lambda_x, \lambda_y \geq 0$, $\xi_x, \xi_y \in \mathbb{R}$ such that

$$\begin{aligned} \langle K_x(x', \widehat{y}) - K_x(\widehat{x}, \widehat{y}), x - \widehat{x}\rangle + \xi_x\|x - \widehat{x}\|^2 \\ \geq \theta_x\|K_y(\widehat{x}, y) - K_y(x, y) - K_{yx}(x, y)(\widehat{x} - x)\| - \frac{\lambda_x}{2}\|x - x'\|^2, \end{aligned} \tag{3.4a}$$

$$\begin{aligned} \langle K_y(x, y) - K_y(x, y') + K_y(\widehat{x}, \widehat{y}) - K_y(\widehat{x}, y), y - \widehat{y}\rangle + \xi_y\|y - \widehat{y}\|^2 \\ \geq \theta_y\|K_x(x', \widehat{y}) - K_x(x', y') - K_{xy}(x', y')(\widehat{y} - y')\| - \frac{\lambda_y}{2}\|y - y'\|^2. \end{aligned} \tag{3.4b}$$

We again elaborate on this assumption. Assumption 3.2 (i)–(iii) are standard in nonlinear optimization of smooth functions. Apart from the estimates in Assumption 3.2 (ii), we make use of the following inequality that is an immediate consequence:

$$\|K_y(x', y) - K_y(x, y) - K_{yx}(x, y)(x' - x)\| \leq \frac{L_{yx}}{2}\|x - x'\|^2. \tag{3.5}$$

The constants ξ_x and ξ_y in Assumption 3.2 (iv) can typically be taken positive by exploiting the strong monotonicity factors γ_G and γ_{F^*} of ∂G and ∂F^*. Indeed, further on in Theorem 4.1, we will require that $\gamma_G - \widetilde{\gamma}_G \geq \xi_x$ and $\gamma_{F^*} - \widetilde{\gamma}_{F^*} \geq \xi_y$, where $\widetilde{\gamma}_G$ and $\widetilde{\gamma}_{F^*}$ will be acceleration factors employed to update the steplength parameters τ_i, ω_i, and σ_i in the algorithm.

In Appendix A we demonstrate that Assumption 3.2 (iv) is closely related to standard second-order optimality conditions, i.e., a positive definite Hessian at the solution $\widehat{u}$. In particular, if the primal problem for the saddle-point functional is strongly convex and the dual problem is strongly concave, the constants that ensure Assumption 3.2 (iv) can be found explicitly. Nonetheless, Assumption 3.2 (iv) is more general than the simple strong convex-concavity. Indeed, in Appendix C we verify Assumption 3.2 for K arising from combinations of a linear operator with a generalized conjugate representations of the step function and the ℓ^0 function from Section 2.2.

Since (3.4b) holds for any $\xi_y, \lambda_y \geq 0$ when $K(x, y) = \langle A(x), y\rangle$ for some $A \in C^1(X)$, the conditions (3.4) reduce to the three-point condition for A from [10] with the exponent $p = 1$. In the present work, such an exponent would correspond to exponents $p_x, p_y \in [1, 2]$ over the norms with the factors θ_y and θ_y that we consider in Assumption B.1 (iv*). These can sometimes be useful: The exponent $p = 2$ was needed in [35, Appendix B] to show the three-point condition for A for a phase and amplitude reconstruction problem. For the sake of readability, in the main part of the present work we focus on the case $p_x = p_y = 1$, i.e., Assumption 3.2 (iv), and discuss the changes needed for $p_x, p_y \in (1, 2]$ in Appendix B.

4 AN ABSTRACT CONVERGENCE RESULT

We want to find a critical point $\widehat{u} = (\widehat{x}, \widehat{y}) \in X \times Y$ of the saddle point functional $(x, y) \mapsto G(x) + K(x, y) - F^*(y)$, i.e., satisfying

$$0 \in H(\widehat{u}) \quad \text{for} \quad H(u) := \begin{pmatrix} \partial G(x) + K_x(x, y) \\ \partial F^*(y) - K_y(x, y) \end{pmatrix}. \tag{4.1}$$

Since G and F^* are proper, convex, and lower semicontinuous, and K is continuously differentiable, using the definition of the saddle-point, the Fréchet derivative, and the convex subdifferential, an elementary limiting argument as in, e.g., [9, Prop. 2.2] shows that the inclusion (4.1) is a first-order necessary optimality condition for a saddle point. If $K(x, y) = \langle Ax, y\rangle$ for $A \in \mathbb{L}(X; Y)$, (4.1) reduces to $-A^*\widehat{y} \in \partial G(\widehat{x})$ and $A\widehat{x} \in \partial F^*(\widehat{y})$, which coincides with the well-known Fenchel–Rockafellar extremality conditions for (1.2); see [14, Remark 4.2].

To study Algorithm 1.1, we reformulate it in the preconditioned proximal point and testing framework of [36]. Specifically, we write Algorithm 1.1 in *implicit proximal point* form as solving in each iteration for $u^{i+1} = (x^{i+1}, y^{i+1}) \in X \times Y$ in

$$0 \in W_{i+1}\widetilde{H}_{i+1}(u^{i+1}) + M_{i+1}(u^{i+1} - u^i), \tag{IPP}$$

where the linearization $\widetilde{H}_{i+1}$ of H, the linear preconditioner M_{i+1}, and the steplength operator W_{i+1} are defined as

$$\widetilde{H}_{i+1}(u) := \begin{pmatrix} \partial G(x) + K_x(x^i, y^i) + K_{xy}(x^i, y^i)(y - y^i) \\ \partial F^*(y) - K_y((1 + \omega_i)x - \omega_i x^i, y^i) - K_{yx}(x^i, y^i)(x - [(1 + \omega_i)x - \omega_i x^i]) \end{pmatrix}, \tag{4.2}$$

$$M_{i+1} := \begin{pmatrix} I & -\tau_i K_{xy}(x^i, y^i) \\ -\omega_i\sigma_{i+1}K_{yx}(x^i, y^i) & I \end{pmatrix}, \tag{4.3}$$

$$W_{i+1} := \begin{pmatrix} \tau_i I & 0 \\ 0 & \sigma_{i+1} I \end{pmatrix}. \tag{4.4}$$

Inserting these definitions into (IPP) and rearranging, we can rewrite inclusion (IPP) as

$$0 \in \begin{pmatrix} \tau_i\partial G(x^{i+1}) + \tau_i K_x(x^i, y^i) + x^{i+1} - x^i \\ \sigma_{i+1}\partial F^*(y^{i+1}) - \sigma_{i+1}K_y((1 + \omega_i)x^{i+1} - \omega_i x^i, y^i) + y^{i+1} - y^i \end{pmatrix}. \tag{4.5}$$

Therefore, based on the definitions of the proximal point mapping $\operatorname{prox}_{\tau G}(v) = (I + \tau\partial G)^{-1}(v)$ and of $\bar{x}^{i+1} = (1 + \omega_i)x^{i+1} - \omega_i x^i$, solving (IPP) for u^{i+1} is equivalent to performing one step of Algorithm 1.1. Since proximal mappings of proper, convex and lower semicontinuous functionals are well-defined, single-valued, and Lipschitz continuous [4, Proposition 12.15], and K is twice Fréchet differentiable on $X \times Y$, this also shows that (IPP) always admits a unique solution u^{i+1}.

The next step is to "test" the inclusion (IPP) by application of $\langle \cdot, u^{i+1} - \widehat{u}\rangle_{Z_{i+1}}$ for the testing operator

$$Z_{i+1} := \begin{pmatrix} \phi_i I & 0 \\ 0 & \psi_{i+1} I \end{pmatrix}.$$

This testing operator and the respective primal and dual testing variables ϕ_i and ψ_{i+1} will be seen to encode convergence rates after some rearrangements of the tested inclusions for $i = 0, \dots, N - 1$.

We will base our convergence analysis on the following abstract estimate, where $\|\cdot\|^2_{Z_{N+1}M_{N+1}}$ forms a local metric that measures the convergence of the iterates while Δ_{i+1} can potentially be used to measure function value or gap converge. In particular, we therefore want $\|u\|_{Z_{N+1}M_{N+1}} \to \infty$ as $N \to \infty$ with as high a rate as possible such that boundedness of $\|u^N - \widehat{u}\|_{Z_{N+1}M_{N+1}}$ implies the convergence of $u^N \to \widehat{u}$ at the best possible rate.

Theorem 4.1 ([36, Theorem 2.1]). *Suppose (IPP) is solvable, and denote the iterates by $\{u^i\}_{i\in\mathbb{N}}$. If $Z_{i+1}M_{i+1}$ is self-adjoint and for some $\widehat{u} \in U$ and $\Delta_{i+1} = \Delta_{i+1}(\widehat{u}) \in \mathbb{R}$, for all $i \leq N-1$,*

$$\langle Z_{i+1}W_{i+1}\widetilde{H}_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle + \Delta_{i+1} \geq \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+2}M_{i+2} - Z_{i+1}M_{i+1}} - \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}, \tag{4.6}$$

then

$$\frac{1}{2}\|u^N - \widehat{u}\|^2_{Z_{N+1}M_{N+1}} \leq \frac{1}{2}\|u^0 - \widehat{u}\|^2_{Z_1M_1} + \sum_{i=0}^{N-1}\Delta_{i+1}. \tag{4.7}$$

The next theorem specializes Theorem 4.1 to our specific setup, converting the abstract condition (4.6) into several steplength and testing parameter update rules and bounds. Specifically, (4.8a) below couples the primal and dual steplengths τ_i and σ_i and the over-relaxation parameter ω_i with the testing parameters. Condition (4.8b) determines convergence rates by limiting how fast the testing parameters can grow. This rate is limited through the available strong monotonicity or second-order behavior ($\gamma_G - \xi_x$ and $\gamma_{F^*} - \xi_y$) through (4.8d) and (4.8e) as well as additional steplength bounds from (4.8c). We point out that only the latter are specific to our non-convex setting; the remaining conditions are present in the convex setting as well, see [36]. We will further develop these rules and conditions in the next section to obtain specific convergence results; an explicit example for a set of parameters satisfying these rules and conditions will be provided for the ℓ^0-TV denoising in Section 7.2 and Appendix C. Here and in the following, we use the notation $\bar{x}^{i+1} := x^{i+1} + \omega_i(x^{i+1} - x^i)$ from Algorithm 1.1 for brevity.

Theorem 4.2. *Suppose Assumptions 3.1 and 3.2 hold with the constants $\theta_x, \theta_y > 0$; $\xi_x, \xi_y \in \mathbb{R}$; $\lambda_x, \lambda_y \geq 0$; $L_{yx} \geq 0$ and $R_K > 0$. For all $i \in \mathbb{N}$, let $\bar{u}^{i+1} := (\bar{x}^{i+1}, y^i)$, and suppose $u^i, u^{i+1}, \widehat{u}, \bar{u}^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$ for some $\rho_x, \rho_y \geq 0$. Assume for all $i \in \mathbb{N}$ that $\overline{\omega} \geq \omega_i \geq \underline{\omega} > 0$ and that for some $0 < \delta \leq \mu < 1$; $\eta_i > 0$; and $\widetilde{\gamma}_G, \widetilde{\gamma}_{F^*} \geq 0$,*

$$\omega_i = \eta_i\eta_{i+1}^{-1}, \qquad \eta_i = \psi_i\sigma_i = \phi_i\tau_i, \tag{4.8a}$$

$$\phi_{i+1} = \phi_i(1 + 2\tau_i\widetilde{\gamma}_G), \qquad \psi_{i+2} = \psi_{i+1}(1 + 2\sigma_{i+1}\widetilde{\gamma}_{F^*}), \tag{4.8b}$$

$$1 \geq \sigma_i\left(\frac{R_K^2\tau_i}{1-\mu} + \frac{\lambda_y}{\omega_i}\right), \qquad \tau_i \leq \frac{\delta}{\lambda_x + L_{yx}(\omega_i + 2)\rho_y}, \tag{4.8c}$$

$$\gamma_G \geq \widetilde{\gamma}_G + \xi_x, \qquad \theta_y \geq \overline{\omega}\rho_x, \tag{4.8d}$$

$$\gamma_{F^*} \geq \widetilde{\gamma}_{F^*} + \xi_y, \qquad \theta_x \geq \rho_y\underline{\omega}^{-1}. \tag{4.8e}$$

Then (4.6) is satisfied for any $\Delta_{i+1} \leq 0$.

Proof. We split the proof into several steps.

Step 1 (estimation of $Z_{i+1}M_{i+1}$) By (4.8a), $\phi_i\tau_i = \eta_i$ and $\psi_{i+1}\sigma_{i+1}\omega_i = \eta_i$, so (4.3) yields

$$Z_{i+1}M_{i+1} = \begin{pmatrix} \phi_i I & -\eta_i K_{xy}(x^i, y^i) \\ -\eta_i K_{yx}(x^i, y^i) & \psi_{i+1} I \end{pmatrix}, \tag{4.9}$$

which is clearly self-adjoint. Applying Cauchy's and Young's inequalities, we further obtain for any $\delta > 0$, $x \in X$, and $y \in Y$ that

$$-2\langle x, \eta_i K_{xy}(x^i, y^i) y\rangle \geq -(1-\delta)\phi_i \|x\|^2 - (1-\delta)^{-1}\phi_i^{-1}\eta_i^2 \|K_{xy}(x^i, y^i) y\|^2,$$

implying that

$$Z_{i+1}M_{i+1} \geq \hat{Q}_{i+1} := \begin{pmatrix} \delta\phi_i I & 0 \\ 0 & \psi_{i+1} I - \frac{\eta_i^2 \phi_i^{-1}}{1-\delta} K_{yx}(x^i, y^i) K_{xy}(x^i, y^i) \end{pmatrix}. \tag{4.10}$$

Step 2 (estimation of $Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}$) Expanding $Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}$ according to (4.9) and then applying (4.8b), we obtain

$$\begin{aligned} \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}} &= -\eta_i \widetilde{\gamma}_G \|x^{i+1} - \widehat{x}\|^2 - \eta_{i+1}\widetilde{\gamma}_{F^*}\|y^{i+1} - \widehat{y}\|^2 \\ &\quad + \langle (\eta_{i+1} K_{xy}(x^{i+1}, y^{i+1}) - \eta_i K_{xy}(x^i, y^i))(y^{i+1} - \widehat{y}), x^{i+1} - \widehat{x}\rangle. \end{aligned} \tag{4.11}$$

Step 3 (estimation of $\widetilde{H}_{i+1}(u^{i+1})$) By (4.2) we have

$$\widetilde{H}_{i+1}(u^{i+1}) = \begin{pmatrix} \partial G(x^{i+1}) + K_x(x^i, y^i) + K_{xy}(x^i, y^i)(y^{i+1} - y^i) \\ \partial F^*(y^{i+1}) - K_y(\bar{x}^{i+1}, y^i) - K_{yx}(x^i, y^i)(x^{i+1} - \bar{x}^{i+1}) \end{pmatrix}.$$

Since $0 \in H(\widehat{u})$, we have $-K_x(\widehat{x}, \widehat{y}) \in \partial G(\widehat{x})$ and $K_y(\widehat{x}, \widehat{y}) \in \partial F^*(\widehat{y})$. Using (4.5) multliplied by Z_{i+1}, Assumption 3.1, and (4.8a), we can thus estimate

$$\begin{aligned} &\langle \widetilde{H}_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{W_{i+1}Z_{i+1}} \\ &\quad \geq \eta_i \gamma_G \|x^{i+1} - \widehat{x}\|^2 + \eta_{i+1}\gamma_{F^*}\|y^{i+1} - \widehat{y}\|^2 \\ &\qquad + \eta_i \langle K_x(x^i, y^i) - K_x(\widehat{x}, \widehat{y}) + K_{xy}(x^i, y^i)(y^{i+1} - y^i), x^{i+1} - \widehat{x}\rangle \\ &\qquad + \eta_{i+1}\langle K_y(\widehat{x}, \widehat{y}) - K_y(\bar{x}^{i+1}, y^i) - K_{yx}(x^i, y^i)(x^{i+1} - \bar{x}^{i+1}), y^{i+1} - \widehat{y}\rangle. \end{aligned} \tag{4.12}$$

Combining (4.12), (4.11), and (4.10), we arrive at

$$\begin{aligned} S_{i+1} &:= \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}} \\ &\quad + \langle \widetilde{H}_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{W_{i+1}Z_{i+1}} \geq \frac{1}{2}\|u^{i+1} - u^i\|^2_{\hat{Q}_{i+1}} + D \end{aligned} \tag{4.13}$$

for

$$\begin{aligned} D &:= \eta_i(\gamma_G - \widetilde{\gamma}_G)\|x^{i+1} - \widehat{x}\|^2 + \eta_{i+1}(\gamma_{F^*} - \widetilde{\gamma}_{F^*})\|y^{i+1} - \widehat{y}\|^2 \\ &\quad + \langle (\eta_{i+1} K_{xy}(x^{i+1}, y^{i+1}) - \eta_i K_{xy}(x^i, y^i))(y^{i+1} - \widehat{y}), x^{i+1} - \widehat{x}\rangle \\ &\quad + \eta_i \langle K_x(x^i, y^i) - K_x(\widehat{x}, \widehat{y}) + K_{xy}(x^i, y^i)(y^{i+1} - y^i), x^{i+1} - \widehat{x}\rangle \\ &\quad + \eta_{i+1}\langle K_y(\widehat{x}, \widehat{y}) - K_y(\bar{x}^{i+1}, y^i) - K_{yx}(x^i, y^i)(x^{i+1} - \bar{x}^{i+1}), y^{i+1} - \widehat{y}\rangle. \end{aligned}$$

The claim of the theorem is established if we prove that $S_{i+1} \geq 0$.

Step 4 (estimation of D) With

$$\begin{aligned}\widetilde{D}_{x+y} := {} & \langle(\eta_{i+1}K_{xy}(x^{i+1},y^{i+1}) - \eta_i K_{xy}(x^i,y^i))(y^{i+1}-\widehat{y}), x^{i+1}-\widehat{x}\rangle \\ & + \eta_i\langle K_x(x^i,y^i) - K_x(\widehat{x},\widehat{y}) + K_{xy}(x^i,y^i)(y^{i+1}-y^i), x^{i+1}-\widehat{x}\rangle \\ & + \eta_{i+1}\langle K_y(\widehat{x},\widehat{y}) - K_y(x^{i+1},y^i), y^{i+1}-\widehat{y}\rangle,\end{aligned}$$

and

$$\begin{aligned}D_\omega := {} & \langle K_y(x^{i+1},y^i) - K_y(\bar{x}^{i+1},y^i) + K_{yx}(x^{i+1},y^i)(\bar{x}^{i+1}-x^{i+1}), y^{i+1}-\widehat{y}\rangle \\ & + \langle[K_{yx}(x^i,y^i) - K_{yx}(x^{i+1},y^i)](\bar{x}^{i+1}-x^{i+1}), y^{i+1}-\widehat{y}\rangle,\end{aligned}$$

we can rewrite

$$D = \eta_i(\gamma_G - \widetilde{\gamma}_G)\|x^{i+1}-\widehat{x}\|^2 + \eta_{i+1}(\gamma_{F^*} - \widetilde{\gamma}_{F^*})\|y^{i+1}-\widehat{y}\|^2 + \widetilde{D}_{x+y} + \eta_{i+1}D_\omega.$$

We rearrange

$$\begin{aligned}\widetilde{D}_{x+y} = {} & \eta_i\langle K_x(x^i,\widehat{y}) - K_x(\widehat{x},\widehat{y}), x^{i+1}-\widehat{x}\rangle \\ & + \eta_{i+1}\langle K_y(\widehat{x},y^{i+1}) - K_y(x^{i+1},y^{i+1}) + K_{yx}(x^{i+1},y^{i+1})(x^{i+1}-\widehat{x}), y^{i+1}-\widehat{y}\rangle \\ & + \eta_{i+1}\langle K_y(\widehat{x},\widehat{y}) - K_y(\widehat{x},y^{i+1}), y^{i+1}-\widehat{y}\rangle \\ & + \eta_{i+1}\langle K_y(x^{i+1},y^{i+1}) - K_y(x^{i+1},y^i), y^{i+1}-\widehat{y}\rangle \\ & - \eta_i\langle K_x(x^i,\widehat{y}) - K_x(x^i,y^i) - K_{xy}(x^i,y^i)(\widehat{y}-y^i), x^{i+1}-\widehat{x}\rangle.\end{aligned}$$

Since $\eta_{i+1} = \eta_i\omega_i^{-1}$, setting

$$\begin{aligned}D_x := {} & \xi_x\|x^{i+1}-\widehat{x}\|^2 + \langle K_x(x^i,\widehat{y}) - K_x(\widehat{x},\widehat{y}), x^{i+1}-\widehat{x}\rangle \\ & + \langle K_y(\widehat{x},y^{i+1}) - K_y(x^{i+1},y^{i+1}) - K_{yx}(x^{i+1},y^{i+1})(\widehat{x}-x^{i+1}), y^{i+1}-\widehat{y}\rangle\omega_i^{-1}, \quad \text{and} \\ D_y := {} & \xi_y\|y^{i+1}-\widehat{y}\|^2 + \langle K_y(\widehat{x},\widehat{y}) - K_y(\widehat{x},y^{i+1}), y^{i+1}-\widehat{y}\rangle \\ & + \langle K_y(x^{i+1},y^{i+1}) - K_y(x^{i+1},y^i), y^{i+1}-\widehat{y}\rangle \\ & - \omega_i\langle K_x(x^i,\widehat{y}) - K_x(x^i,y^i) - K_{xy}(x^i,y^i)(\widehat{y}-y^i), x^{i+1}-\widehat{x}\rangle,\end{aligned}$$

we can write

$$\begin{aligned}D = {} & \eta_i(\gamma_G - \widetilde{\gamma}_G - \xi_y)\|x^{i+1}-\widehat{x}\|^2 + \eta_{i+1}(\gamma_{F^*} - \widetilde{\gamma}_{F^*} + \xi_x)\|y^{i+1}-\widehat{y}\|^2 \\ & + \eta_i D_x + \eta_{i+1}D_y + \eta_{i+1}D_\omega.\end{aligned}$$

As for the estimate for D_ω, using Assumption 3.2 (ii) and (3.5) we obtain

$$\begin{aligned}D_\omega &\geq -\frac{L_{yx}}{2}\|\bar{x}^{i+1}-x^{i+1}\|^2\|y^{i+1}-\widehat{y}\| - L_{yx}\|x^{i+1}-x^i\|\|\bar{x}^{i+1}-x^{i+1}\|\|y^{i+1}-\widehat{y}\| \\ &\geq -\frac{L_{yx}\omega_i(\omega_i+2)\rho_y}{2}\|x^{i+1}-x^i\|^2\end{aligned} \tag{4.14}$$

using in the last inequality the expansion $\bar{x}^{i+1} := x^{i+1} + \omega_i(x^{i+1}-x^i)$ and the bound $\|y^{i+1}-\widehat{y}\| \leq \rho_y$ that follows from the assumed inclusion $u^{i+1} \in \mathcal{U}(\rho_x,\rho_y)$.

We now use Assumption 3.2 (iv) to further bound D_x and D_y. From (3.4a), we obtain

$$\begin{aligned}
D_x &\geq \theta_x \|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\| - \frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 \\
&\quad - \|y^{i+1} - \widehat{y}\|\|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\|\omega_i^{-1} \\
&\geq (\theta_x - \rho_y \underline{\omega}^{-1})\|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\| \\
&\quad - \frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 \\
&\geq -\frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2,
\end{aligned} \tag{4.15}$$

using in the last two inequalities that $u^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$ for some $\rho_x, \rho_y \geq 0$, $\omega_i^{-1} \leq \underline{\omega}^{-1}$ and $\theta_x \geq \rho_y \underline{\omega}^{-1}$ from (4.8e). Analogously, from (3.4b) and Cauchy's inequality,

$$\begin{aligned}
D_y &\geq \theta_y \|K_x(x^i, \widehat{y}) - K_x(x^i, y^i) - K_{xy}(x^i, y^i)(\widehat{y} - y^i)\| - \frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2 \\
&\quad - \omega_i \|x^{i+1} - \widehat{x}\|\|K_x(x^i, \widehat{y}) - K_x(x^i, y^i) - K_{xy}(x^i, y^i)(\widehat{y} - y^i)\| \\
&\geq (\theta_y - \rho_x \overline{\omega})\|K_x(x^i, \widehat{y}) - K_x(x^i, y^i) - K_{xy}(x^i, y^i)(\widehat{y} - y^i)\| \\
&\quad - \frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2 \\
&\geq -\frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2,
\end{aligned} \tag{4.16}$$

where in the last two inequalities we again used $u^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$, $\omega_i \leq \overline{\omega}$, and $\theta_y \geq \overline{\omega}\rho_x$ from (4.8d). Therefore, combining (4.14), (4.15), and (4.16), we obtain

$$\begin{aligned}
D &= \eta_i D_x + \eta_{i+1} D_y + \eta_{i+1} D_\omega + \eta_i(\gamma_G - \widetilde{\gamma}_G - \xi_x)\|x^{i+1} - \widehat{x}\|^2 \\
&\quad + \eta_{i+1}(\gamma_{F^*} - \widetilde{\gamma}_{F^*} - \xi_y)\|y^{i+1} - \widehat{y}\|^2 \\
&\geq \eta_{i+1}(\gamma_{F^*} - \widetilde{\gamma}_{F^*} - \xi_y)\|y^{i+1} - \widehat{y}\|^2 - \eta_i \frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 \\
&\quad + \eta_i(\gamma_G - \widetilde{\gamma}_G - \xi_x)\|x^{i+1} - \widehat{x}\|^2 - \eta_{i+1}\frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2 \\
&\quad - \eta_i \frac{L_{yx}}{2}(\omega_i + 2)\rho_y \|x^{i+1} - x^i\|^2 \\
&\geq -\eta_i \frac{\lambda_x + L_{yx}(\omega_i + 2)\rho_y}{2}\|x^{i+1} - x^i\|^2 - \eta_{i+1}\frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2,
\end{aligned} \tag{4.17}$$

where we have also used the first bounds of (4.8d) and (4.8e) in the final step. Further using (4.8c) and $\eta_{i+1} = \eta_i \omega_i^{-1}$, we deduce that $D \geq -\frac{1}{2}\|u^{i+1} - u^i\|^2_{\hat{Q}_{i+1}}$. Recalling (4.13), we obtain $S_{i+1} \geq 0$, i.e., (4.6) holds with $\Delta_{i+1} \leq 0$ as claimed. □

In the subsequent sections, we will also need the following corollary.

Corollary 4.3. *Suppose that Assumption 3.2 (iii) and the conditions (4.8) hold. Then*

$$(1 - \mu)\psi_{i+1} \geq \eta_i^2 \phi_i^{-1} R_K^2$$

and

$$Z_{i+1}M_{i+1} \geq \begin{pmatrix} \delta\phi_i I & 0 \\ 0 & (\mu-\delta)(1-\delta)^{-1}\psi_{i+1}I \end{pmatrix}. \tag{4.18}$$

Proof. Observe that due to (4.8),

$$(1-\mu)\psi_{i+1} \geq (1-\mu)\psi_i = \frac{(1-\mu)\eta_i^2}{\sigma_i\tau_i\phi_i} \geq \eta_i^2\phi_i^{-1}R_K^2.$$

This is our first claim. As for the second term, from Assumption 3.2 (iii) we have

$$\frac{\eta_i^2\phi_i^{-1}}{1-\delta}K_{yx}(x^i,y^i)K_{xy}(x^i,y^i) \leq \frac{\eta_i^2\phi_i^{-1}}{1-\delta}R_K^2 I \leq \frac{1-\mu}{1-\delta}\psi_{i+1}I.$$

Inserting this bound into (4.10) in the proof of Theorem 4.2 establishes (4.18). □

5 LOCAL STEPLENGTH BOUNDS

In the previous section, we derived steplength conditions that we will further develop in Section 6 to prove convergence and convergence rates. However, we implicitly required that all the iterations $\{u^i\}_{i\in\mathbb{N}}$ belong to $\mathcal{U}(\rho_x,\rho_y)$. In this section, we derive additional steplengths restrictions to ensure that this holds.

We start with a lemma that bounds the next iterate u^{i+1} given bounds on the current iterate u^i and the steplengths for the current iteration. Afterwards, we chain these estimates to only require bounds on the initial iterates and the steplengths.

Lemma 5.1. *Fix $i \in \mathbb{N}$. Suppose Assumption 3.1, Assumption 3.2 (ii), and (iii) hold in $\mathcal{U}(\rho_x,\rho_y)$, and that u^{i+1} solves (IPP). For simplicity, assume $\omega_i \leq 1$. Suppose $r_{x,i}, r_{y,i}, \delta_x, \delta_y > 0$ and $\widehat{u} \in H^{-1}(0)$ are such that $\mathbb{B}(\widehat{x}, r_{x,i}+\delta_x) \times \mathbb{B}(\widehat{y}, r_{y,i}+\delta_y) \subseteq \mathcal{U}(\rho_x,\rho_y)$ and $u^i \in \mathbb{B}(\widehat{x}, r_{x,i}) \times \mathbb{B}(\widehat{y}, r_{y,i})$. If*

$$\tau_i \leq \frac{\delta_x}{2R_K r_{y,i} + 2L_x(\widehat{y})r_{x,i}} \quad \text{and} \quad \sigma_{i+1} \leq \frac{\delta_y}{L_y(\widehat{x})r_{y,i} + R_K(r_{x,i}+\delta_x)}, \tag{5.1}$$

then $u^{i+1} \in \mathbb{B}(\widehat{x}, r_{x,i}+\delta_x) \times \mathbb{B}(\widehat{y}, r_{y,i}+\delta_y)$ and $\|\bar{x}^{i+1} - \widehat{x}\| \leq r_{x,i} + \delta_x$.

Proof. We want to show that the steplength conditions (5.1) are sufficient for

$$\|x^{i+1} - \widehat{x}\| \leq r_{x,i} + \delta_x, \quad \|\bar{x}^{i+1} - \widehat{x}\| \leq r_{x,i} + \delta_x, \quad \text{and} \quad \|y^{i+1} - \widehat{y}\| \leq r_{y,i} + \delta_y.$$

We do this by applying the testing argument on the primal and dual variables separately. Multiplying (IPP) by $Z_{i+1}^*(u^{i+1} - \widehat{u})$ with $\phi_i = 1$ and $\psi_{i+1} = 0$, we obtain

$$0 \in \tau_i\langle \partial G(x^{i+1}) + K_x(x^i,y^i), x^{i+1} - \widehat{x}\rangle + \langle x^{i+1} - x^i, x^{i+1} - \widehat{x}\rangle.$$

Using the three-point identity

$$\langle x^{i+1} - x^i, x^{i+1} - \widehat{x}\rangle = \frac{1}{2}\|x^{i+1} - x^i\|^2 - \frac{1}{2}\|x^i - \widehat{x}\|^2 + \frac{1}{2}\|x^{i+1} - \widehat{x}\|^2, \tag{5.2}$$

we obtain

$$\|x^i - \widehat{x}\|^2 \in 2\tau_i \langle \partial G(x^{i+1}) + K_x(x^i, y^i), x^{i+1} - \widehat{x}\rangle + \|x^{i+1} - x^i\|^2 + \|x^{i+1} - \widehat{x}\|^2.$$

Using further $0 \in \partial G(\widehat{x}) + K_x(\widehat{x}, \widehat{y})$ and the monotonicity of ∂G, we arrive at

$$\|x^{i+1} - x^i\|^2 + \|x^{i+1} - \widehat{x}\|^2 + 2\tau_i \langle K_x(x^i, y^i) - K_x(\widehat{x}, \widehat{y}), x^{i+1} - \widehat{x}\rangle \leq \|x^i - \widehat{x}\|^2.$$

With $C_x := \tau_i \|K_x(x^i, y^i) - K_x(\widehat{x}, \widehat{y})\|$, this implies that

$$\|x^{i+1} - x^i\|^2 + \|x^{i+1} - \widehat{x}\|^2 \leq 2C_x \|x^{i+1} - \widehat{x}\| + \|x^i - \widehat{x}\|^2. \tag{5.3}$$

After rearranging the terms and using $\|x^{i+1} - \widehat{x}\| \leq \|x^{i+1} - x^i\| + \|x^i - \widehat{x}\|$, we thus have

$$(\|x^{i+1} - x^i\| - C_x)^2 + \|x^{i+1} - \widehat{x}\|^2 \leq (\|x^i - \widehat{x}\| + C_x)^2,$$

which leads to

$$\|x^{i+1} - \widehat{x}\| \leq \|x^i - \widehat{x}\| + C_x. \tag{5.4}$$

To estimate the dual variable, we multiply (IPP) by $Z_{i+1}^*(u^{i+1} - \widehat{u})$ with $\phi_i = 0$ and $\psi_{i+1} = 1$. This gives

$$0 \in \sigma_{i+1} \langle \partial F^*(y^{i+1}) - K_y(\bar{x}^{i+1}, y^i), y^{i+1} - \widehat{y}\rangle + \langle y^{i+1} - y^i, y^{i+1} - \widehat{y}\rangle.$$

Using $0 \in \partial F^*(\widehat{y}) - K_y(\widehat{x}, \widehat{y})$ and following the steps leading to (5.4), we deduce

$$\|y^{i+1} - \widehat{y}\| \leq \|y^i - \widehat{y}\| + C_y \tag{5.5}$$

with $C_y := \sigma_{i+1} \|K_y(\widehat{x}, \widehat{y}) - K_y(\bar{x}^{i+1}, y^i)\|$.

We now proceed to derive bounds on C_x and C_y with the goal of bounding both (5.4) and (5.5) from above. Using Assumption 3.2 (ii), (iii), and the mean value theorem applied to $K_x(x^i, \cdot)$ and $K_y(\cdot, y^i)$,

$$\begin{aligned}
C_x &\leq \tau_i (\|K_x(x^i, y^i) - K_x(x^i, \widehat{y})\| + \|K_x(x^i, \widehat{y}) - K_x(\widehat{x}, \widehat{y})\|) \\
&\leq \tau_i (R_K r_{y,i} + L_x(\widehat{y}) r_{x,i}) =: R_x, \\
C_y &\leq \sigma_{i+1} (\|K_y(\widehat{x}, \widehat{y}) - K_y(\widehat{x}, y^i)\| + \|K_y(\widehat{x}, y^i) - K_y(\bar{x}^{i+1}, y^i)\| \\
&\leq \sigma_{i+1} (L_y(\widehat{x}) r_{y,i} + R_K(r_{x,i} + \delta_x)) =: R_y,
\end{aligned}$$

the latter under the assumption that $\|\bar{x}^{i+1} - \widehat{x}\| \leq r_{x,i} + \delta_x$, which we now verify. First, by definition,

$$\begin{aligned}
\|\bar{x}^{i+1} - \widehat{x}\|^2 &= \|x^{i+1} - \widehat{x} + \omega_i (x^{i+1} - x^i)\|^2 \\
&= \|x^{i+1} - \widehat{x}\|^2 + \omega_i^2 \|x^{i+1} - x^i\|^2 + 2\omega_i \langle x^{i+1} - \widehat{x}, x^{i+1} - x^i\rangle \\
&= (1 + \omega_i) \|x^{i+1} - \widehat{x}\|^2 + \omega_i (1 + \omega_i) \|x^{i+1} - x^i\|^2 - \omega_i \|x^i - \widehat{x}\|^2 \\
&\leq (1 + \omega_i)(\|x^{i+1} - \widehat{x}\|^2 + \|x^{i+1} - x^i\|^2) - \omega_i \|x^i - \widehat{x}\|^2.
\end{aligned}$$

Applying (5.3) and (5.4), we obtain

$$\begin{aligned}\|\bar{x}^{i+1}-\widehat{x}\|^2 &\le (1+\omega_i)(2C_x\|x^{i+1}-\widehat{x}\|+\|x^i-\widehat{x}\|^2)-\omega_i\|x^i-\widehat{x}\|^2\\ &\le 4C_x\|x^{i+1}-\widehat{x}\|+\|x^i-\widehat{x}\|^2 \le 4C_x(\|x^i-\widehat{x}\|+C_x)+\|x^i-\widehat{x}\|^2 \le (2C_x+r_{x,i})^2.\end{aligned}$$

The bound (5.1) on τ_i implies that $C_x \le R_x \le \delta_x/2$ and hence that $\|\bar{x}^{i+1}-\widehat{x}\| \le r_{x,i}+\delta_x$. From (5.4) we thus obtain $\|x^{i+1}-\widehat{x}\| \le r_{x,i}+\delta_x$. The bound (5.1) on σ_i then implies that $C_y \le R_y \le \delta_y$, which together with (5.5) completes the proof. □

To chain the applications of Lemma 5.1 on each iteration $i \in \mathbb{N}$, we introduce the following assumption, for which we recall the notations in Assumption 3.2 as well as the definition of $\mathcal{U}(\rho_x,\rho_y)$ from (3.3).

Assumption 5.1. Suppose Assumption 3.2 holds near a solution $\widehat{u} \in H^{-1}(0)$. Given an initial iterate $u^0 \in X \times Y$, and initial steplength parameters $\tau_0, \sigma_1, \omega_0 > 0$ as well as $0 < \delta \le \mu < 1$ (to satisfy (4.8)), define the weighted distance

$$r_{\max} := \sqrt{2\delta^{-1}(\|x^0-\widehat{x}\|^2+\nu^{-1}\|y^0-\widehat{y}\|^2)} \quad \text{with} \quad \nu := \sigma_1\omega_0\tau_0^{-1}. \tag{5.6}$$

We then assume that there exist $\delta_x, \delta_y > 0$ and $r_y \ge r_{\max}\sqrt{\nu(1-\delta)\delta(\mu-\delta)^{-1}}$ such that

$$\mathbb{B}(\widehat{x}, r_{\max}+\delta_x)\times\mathbb{B}(\widehat{y}, r_y+\delta_y) \subseteq \mathcal{U}(\rho_x,\rho_y)$$

and that for all $i \in \mathbb{N}$ the steplengths $\tau_i, \sigma_i > 0$ satisfy

$$\tau_i \le \frac{\delta_x}{2R_K r_y + 2L_x(\widehat{y})r_{\max}} \quad \text{and} \quad \sigma_{i+1} \le \frac{\delta_y}{L_y(\widehat{x})r_y + R_K(r_{\max}+\delta_x)}. \tag{5.7}$$

Lemma 5.2. *For all $i \in \mathbb{N}$, suppose u^{i+1} solves (IPP) and that all the conditions of Theorem 4.2 are satisfied for some $\rho_x, \rho_y > 0$ and $\widetilde{\gamma}_G, \widetilde{\gamma}_{F^*} \ge 0$ except for the requirement $u^i, u^{i+1}, \bar{u}^{i+1} \in \mathcal{U}(\rho_x,\rho_y)$. Then if Assumption 5.1 holds, $\{u^i\}_{i\in\mathbb{N}}, \{\bar{u}^{i+1}\}_{i\in\mathbb{N}} \subset \mathcal{U}(\rho_x,\rho_y)$.*

Proof. We define $r_{x,i} := \frac{1}{\sqrt{\delta\phi_i}}\|u^0-\widehat{u}\|_{Z_1M_1}$ and

$$\mathcal{U}_i := \left\{(x,y) \in X\times Y \;\middle|\; \|x-\widehat{x}\|^2 + \tfrac{\psi_{i+1}}{\phi_i}\tfrac{\mu-\delta}{(1-\delta)\delta}\|y-\widehat{y}\|^2 \le r_{x,i}^2\right\}.$$

Since the conditions (4.8) hold, we can apply Corollary 4.3 and the estimate (4.18) on $Z_{i+1}M_{i+1}$ to deduce that

$$\{u \in X\times Y \mid \|u-\widehat{u}\|_{Z_{i+1}M_{i+1}} \le \|u^0-\widehat{u}\|_{Z_1M_1}\} \subset \mathcal{U}_i. \tag{5.8}$$

From (4.8b), we also deduce that $\phi_{i+1} \ge \phi_i$ and hence that $r_{x,i+1} \le r_{x,i}$. Consequently, if $r_{x,0} \le r_{\max}$, then

$$\mathbb{B}(\widehat{x}, r_{x,i}+\delta_x)\times\mathbb{B}(\widehat{y}, r_y+\delta_y) \subseteq \mathbb{B}(\widehat{x}, r_{\max}+\delta_x)\times\mathbb{B}(\widehat{y}, r_y+\delta_y) \subseteq \mathcal{U}(\rho_x,\rho_y), \tag{5.9}$$

so it will suffice to show that $u^i \in \mathbb{B}(\widehat{x}, r_{x,i}+\delta_x)\times\mathbb{B}(\widehat{y}, r_y+\delta_y)$ for each $i \in \mathbb{N}$ to prove the claim. We do this in two steps. In the first step, we show that $r_{x,i} \le r_{\max}$ and

$$\mathcal{U}_i \subseteq \mathbb{B}(\widehat{x}, r_{x,i})\times\mathbb{B}(\widehat{y}, r_y) \quad (i \in \mathbb{N}). \tag{5.10}$$

In the second step, we show by induction that $u^i \in \mathcal{U}_i$ as well as $\bar{u}^{i+1} \in \mathcal{U}(\rho_x,\rho_y)$ for $i \in \mathbb{N}$.

Step 1 We first prove (5.10). Since $\mathcal{U}_i \subseteq \mathbb{B}(\widehat{x}, r_{x,i}) \times Y$, we only have to show that $\mathcal{U}_i \subseteq X \times \mathbb{B}(\widehat{y}, r_y)$. First, note that (4.8) and $\widetilde{\gamma}_G, \widetilde{\gamma}_{F^*} \geq 0$ imply $\psi_{i+1} \geq \psi_i \geq \psi_1$ as well as $\phi_{i+1} \geq \phi_i \geq \phi_0 = \eta_1 \omega_0 \tau_0^{-1} = \nu\psi_1$ for ν defined in (5.6). We then obtain from the definition of $r_{x,i}$ substituting $Z_1 M_1$ from (4.9) that

$$r_{x,i}^2 \delta\phi_i = \|u^0 - \widehat{u}\|_{Z_1 M_1}^2 = \nu\psi_1 \|x^0 - \widehat{x}\|^2 - 2\eta_0 \langle x^0 - \widehat{x}, K_{xy}(x^0, y^0)(y^0 - \widehat{y})\rangle + \psi_1 \|y^0 - \widehat{y}\|^2.$$

Using Cauchy's and Young's inequalities, the fact that $\phi_i \geq \nu\psi_1$, and the assumption that $\|K_{xy}(x^0, y^0)\| \leq R_K$, we arrive at

$$r_{x,i}^2 \leq (2\nu\psi_1 \|x^0 - \widehat{x}\|^2 + (\psi_1 + \eta_0^2 \phi_0^{-1} R_K^2)\|y^0 - \widehat{y}\|^2)(\delta\nu\psi_1)^{-1}.$$

We obtain from Corollary 4.3 that $\eta_0^2 \phi_0^{-1} R_K^2 \leq (1-\mu)\psi_1 \leq \psi_1$ and hence that $r_{x,i}^2 \leq r_{\max}^2$. The assumption on r_y then yields for all $i \in \mathbb{N}$ that

$$r_y^2 \geq r_{\max}^2 \frac{\phi_0}{\psi_1} \frac{(1-\delta)\delta}{\mu - \delta} \geq \frac{r_{x,0}^2 \phi_0}{\psi_{i+1}} \frac{(1-\delta)\delta}{\mu - \delta} = \frac{r_{x,i}^2 \phi_i}{\psi_{i+1}} \frac{(1-\delta)\delta}{\mu - \delta}. \tag{5.11}$$

Thus (5.10) follows from the definition of $\mathcal{U}_i$.

Step 2 We next show by induction that $u^i \in \mathcal{U}_i$ and $\bar{u}^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$ for all $i \in \mathbb{N}$. Since (5.8) holds for $i = 0$, we have that $u^0 \in \mathcal{U}_0$. Moreover, since in Step 1 we have $r_{x,0} \leq r_{\max}$, the bound (5.1) for $i = 0$ follows from (5.7). This gives the induction basis.

Suppose now that $u^N \in \mathcal{U}_N$. By (5.10), we have that $u^N \in \mathbb{B}(\widehat{x}, r_{x,N}) \times \mathbb{B}(\widehat{y}, r_y)$. Since again the bound (5.1) for $i = N$ follows from (5.7) and the bound $r_{x,N} \leq r_{\max}$ follows from Step 1, we can apply Lemma 5.1 to obtain

$$u^{N+1} \in \mathbb{B}(\widehat{x}, r_{x,N} + \delta_x) \times \mathbb{B}(\widehat{y}, r_y + \delta_y) \quad \text{and} \quad \bar{x}^{N+1} \in \mathbb{B}(\widehat{x}, r_{x,N} + \delta_x).$$

By (5.9), we have $\mathbb{B}(\widehat{x}, r_{x,N} + \delta_x) \times \mathbb{B}(\widehat{y}, r_y + \delta_y) \subseteq \mathcal{U}(\rho_x, \rho_y)$ and thus $u^{N+1}, \bar{u}^{N+1} \in \mathcal{U}(\rho_x, \rho_y)$. Theorem 4.2 now implies that (4.7) is satisfied for $i \leq N$ with $\Delta_{N+1} \leq 0$, which together with (4.7) and (5.8) yields that $u^{N+1} \in \mathcal{U}_{N+1}$. This completes the induction step and hence the proof. □

6 CONVERGENCE ESTIMATES

We are now ready to formulate the main convergence results of this paper based on the estimates derived above. First, based on (4.8d) and (4.8e), strong convexity may be required if ξ_x and ξ_y have to be positive for Assumption 3.2 to be satisfied. Moreover, the neighborhood $\mathcal{U}(\rho_x, \rho_y)$ has to be small enough, as determined by the assumptions $\theta_x \geq \rho_y \underline{\omega}^{-1}$ and $\theta_y \geq \overline{\omega}\rho_x$ in the next results. This affects the admissible steplengths and how close we have to initialize u^0 via Assumption 5.1. After the next three main convergence results, we show that Assumption 5.1 is satisfied if we initialize close enough to a root $\widehat{u} \in H^{-1}(0)$. Hence, to apply the theorems in practice, we have to find constants for which Assumptions 3.1 and 3.2 are satisfied, use these constants to bound and compute the steplengths as described in the theorems, and initialize close enough to $\widehat{u}$. In Appendix B we consider some relaxation of Assumption 3.2 (iv), which in turn requires larger γ_G and γ_{F^*} instead of $\theta_x \geq \rho_y \underline{\omega}^{-1}$ and $\theta_y \geq \overline{\omega}\rho_x$.

The following theorem provides conditions sufficient for weak convergence of the sequence $\{u^i\}_{i\in\mathbb{N}}$ generated by Algorithm 1.1. Apart from technical requirements of Theorem 4.2, we require additional weak-to-strong continuity of the mapping $u \mapsto K_{yx}(u)x$. While its verification depends on the particular choice of K, it is trivially satisfied in two cases: (i) X and Y are finite-dimensional and K_{yx} is continuous; or (ii) the mapping $u \mapsto K_{yx}(u)x$ is linear and compact.

Theorem 6.1 (weak convergence: $\omega_i = 1$). *Suppose Assumptions 3.1, 3.2 and 5.1 hold for some $R_K > 0$; $L_{yx} \geq 0$; $\lambda_x, \lambda_y, \theta_x, \theta_y \geq 0$; and $\xi_x, \xi_y \in \mathbb{R}$ such that*

$$\xi_x = \gamma_G, \qquad \theta_y \geq 2\rho_x, \tag{6.1a}$$

$$\xi_y = \gamma_{F^*}, \qquad \theta_x \geq 2\rho_y. \tag{6.1b}$$

For some $0 < \delta < \mu < 1$, choose

$$\tau_i \equiv \tau < \frac{\delta}{\lambda_x + 3L_{yx}\rho_y}, \qquad \sigma_i \equiv \sigma \leq \left(\frac{R_K^2\tau}{1-\mu} + \lambda_y\right)^{-1}, \quad \text{and} \quad \omega_i \equiv 1. \tag{6.2}$$

Furthermore, suppose that

(i) $u^i \rightharpoonup \bar{u}$ implies that $K_{yx}(u^i)x \to K_{yx}(\bar{u})x$ for all $x \in X$,

and either

(iia) the mapping $u \mapsto (K_x(u), K_y(u))$ is weak-to-strong continuous in $\mathcal{U}(\rho_x, \rho_y)$; or

(iib) the mapping $u \mapsto (K_x(u), K_y(u))$ is weak-to-weak continuous, but Assumption 3.1 (monotone ∂G and ∂F^*) and Assumption 3.2 (iv) (three-point condition on K) hold at any *weak limit $\bar{u} = (\bar{x}, \bar{y})$ of $\{u^i\}_{i\in\mathbb{N}}$ for the same choices of θ_x and θ_y.*

Then the sequence $\{u^i\}_{i\in\mathbb{N}}$ generated by Algorithm 1.1 converges weakly to some $\bar{u} \in H^{-1}(0)$ (possibly different from $\widehat{u}$).

Since it is assumed that $\theta_x \geq 2\rho_y$, we can replace ρ_y by $\theta_x/2$ in the bound on τ in (6.2) if the latter is more readily available.

For constant τ, σ, and $\omega = 1$, we have to set $\psi_i \equiv \psi$ and $\phi_i \equiv \phi$ to satisfy (4.8a). Consequently, applying Corollary 4.3 to bound $Z_{i+1}M_{i+1}$ from below will not help to prove Theorem 6.1. We instead will make use of the following enhanced version of Opial's lemma.

Lemma 6.2 ([10, Lemma A.2]). *Let U be a Hilbert space, $\hat{U} \subset U$ (not necessarily closed or convex), and $\{u^i\}_{i\in\mathbb{N}} \subset U$. Also let $A_i \in \mathbb{L}(U;U)$ be self-adjoint and $A_i \geq \hat{\epsilon}^2 I$ for some $\hat{\epsilon} \neq 0$ for all $i \in \mathbb{N}$. If the following conditions hold, then $u^i \rightharpoonup \bar{u}$ in U for some $\bar{u} \in \hat{U}$:*

(i) The sequence $\{\|u^i - \hat{u}\|_{A_i}\}_{i\in\mathbb{N}}$ is nonincreasing for some *$\hat{u} \in \hat{U}$.*

(ii) All weak limit points of $\{u^i\}_{i\in\mathbb{N}}$ belong to $\hat{U}$.

(iii) There exists $C > 0$ such that $\|A_i\| \leq C^2$ for all i, and for any weakly convergent subsequence $\{u_{i_k}\}_{k\in\mathbb{N}}$ there exists $A_\infty \in \mathbb{L}(U;U)$ such that $A_{i_k}u \to A_\infty u$ strongly in U for all $u \in U$.

Proof of Theorem 6.1. We first verify (4.8) so that we can apply Theorem 4.2 and Lemma 5.2. We set $\psi_N \equiv 1$, $\phi_N \equiv \sigma\tau^{-1}$, $\widetilde{\gamma}_G = \widetilde{\gamma}_{F^*} = 0$ to satisfy (4.8a), (4.8b), (4.8d) and (4.8e) for $\omega = \underline{\omega} = \overline{\omega} = 1$ and ξ_x, ξ_y, θ_x, θ_y satisfying (6.1). With the choice $\omega = 1$, the bounds (6.2) thus ensure (4.8c).

Hence (4.8) holds, which together with Assumption 5.1 and $\psi_1 = 1$ enables us to use Lemma 5.2 to obtain $\{u^i\}_{i\in\mathbb{N}} \in \mathcal{U}(\rho_x, \rho_y)$ and $\{\bar{x}^{i+1}\}_{i\in\mathbb{N}} \in \mathbb{B}(\widehat{x}, \rho_x)$. Therefore there exists at least one weak limit point of $\{u^i\}_{i\in\mathbb{N}}$. Moreover, (4.9) yields self-adjointness of $Z_{i+1}M_{i+1}$ and since the bounds (6.2) are strict, Theorem 4.2 holds with $\Delta_{i+1} \le -\hat{\delta}\sum_{i=0}^{N}\|u^{i+1}-u^i\|^2$ for some $\hat{\delta} > 0$.

We now verify the conditions of Lemma 6.2 with $\hat{U} = H^{-1}(0)$ and $A_i = Z_{i+1}M_{i+1}$. Estimate (4.7) is valid for any starting iterate; thus setting $N = 1$ and taking u^i instead of u^0, we obtain $\|u^{i+1}-\widehat{u}\|^2_{Z_{i+2}M_{i+2}} \le \|u^i-\widehat{u}\|^2_{Z_{i+1}M_{i+1}} + \Delta_{i+1}$ for any $\Delta_{i+1} \le 0$ due to Theorem 4.2. This verifies (i). Moreover, (iii) follows from the assumed constant steplengths, Assumption 3.2 (iii), and the assumption that $K_{yx}(u^i)x \to K_{yx}(\bar{u})x$ for all $x \in X$ if $u^i \rightharpoonup \bar{u}$.

Hence we only need to verify (ii), i.e., if a subsequence of $\{u^i\}_{i\in\mathbb{N}}$ converges weakly to some $\bar{u}$, then $\bar{u} \in H^{-1}(0)$. We note that $W_{i+1} \equiv W$, and (IPP) implies that $v_{i+1} \in WA(u^{i+1})$ for

$$A(u^{i+1}) := \begin{pmatrix} \partial G(x^{i+1}) - \gamma_G(x^{i+1}-\bar{x}) \\ \partial F^*(y^{i+1}) - \gamma_{F^*}(y^{i+1}-\bar{y}) \end{pmatrix} \tag{6.3}$$

$$\begin{aligned} v_{i+1} := W\begin{pmatrix} -K_x(x^{i+1},y^{i+1}) - \gamma_G(x^{i+1}-\bar{x}) \\ K_y(x^{i+1},y^{i+1}) - \gamma_{F^*}(y^{i+1}-\bar{y}) \end{pmatrix} - M_{i+1}(u^{i+1}-u^i) \\ - W\begin{pmatrix} K_x(x^i,y^i) - K_x(x^{i+1},y^{i+1}) + K_{xy}(x^i,y^i)(y^{i+1}-y^i) \\ K_y(x^{i+1},y^{i+1}) - K_y(\bar{x}^{i+1},y^i) - K_{yx}(x^i,y^i)(x^{i+1}-\bar{x}^{i+1}) \end{pmatrix}. \end{aligned} \tag{6.4}$$

Therefore it suffices to show that if $u^{i_k} \rightharpoonup \bar{u} = (\bar{x},\bar{y})$ for a subsequence, then

$$v_{i_k} \rightharpoonup \bar{v} := W\begin{pmatrix} -K_x(\bar{x},\bar{y}) \\ K_y(\bar{x},\bar{y}) \end{pmatrix} \qquad \text{and} \qquad \bar{v} \in WA(\bar{u}),$$

which by construction is equivalent to $\bar{u} \in H^{-1}(0)$. Note that A is maximally monotone since it only involves subgradient mappings of proper convex lower semicontinuous functions due to Assumption 3.1. Moreover, further use of (4.7) shows that $\sum_{i=0}^{\infty}\frac{\hat{\delta}}{2}\|u^{i+1}-u^i\|^2 < \infty$ and hence that $\|u^{i+1}-u^i\| \to 0$. The last two terms in (6.4) thus converge strongly to zero. We therefore only have to consider the first term, for which we make a case distinction.

(a) If assumption (iia) holds, we obtain that $v_{i_k} \to \bar{v}$, and the required inclusion $\bar{v} \in A(\bar{u})$ follows from the fact that the graph of the maximally monotone operator A is sequentially weakly–strongly closed; see [4, Proposition 16.36].

(b) If assumption (iib) holds, then only $v_{i_k} \rightharpoonup \bar{v}$. In this case, we can apply the Brezis–Crandall–Pazy Lemma [4, Corollary 20.59 (iii)] to obtain the required inclusion under the additional condition that $\limsup_{k\to\infty}\langle u^{i_k}-\bar{u}, v_{i_k}-\bar{v}\rangle \le 0$. In our case, recalling that the last two terms of (6.4) converge strongly to zero, we have that

$$\limsup_{k\to\infty}\langle u^{i_k}-\bar{u}, v_{i_k}-\bar{v}\rangle \le \limsup_{i\to\infty}\langle u_i-\bar{u}, v_i-\bar{v}\rangle = \limsup_{i\to\infty} q_i$$

for

$$q_i := \langle K_x(\bar{x},\bar{y}) - K_x(x^{i+1},y^{i+1}), x^{i+1}-\bar{x}\rangle + \langle K_y(x^{i+1},y^{i+1}) - K_y(\bar{x},\bar{y}), y^{i+1}-\bar{y}\rangle \\ - \gamma_{F^*}\|y^{i+1}-\bar{y}\|^2 - \gamma_G\|x^{i+1}-\bar{x}\|^2.$$

Defining

$$\begin{aligned} d_i^x &:= \langle K_y(x^{i+1},y^{i+1}) - K_y(\bar{x},y^{i+1}) + K_{yx}(x^{i+1},y^{i+1})(\bar{x}-x^{i+1}), y^{i+1}-\bar{y}\rangle \\ &\quad - \langle K_x(x^i,\bar{y}) - K_x(\bar{x},\bar{y}), x^{i+1}-\bar{x}\rangle - \gamma_G\|x^{i+1}-\bar{x}\|^2 \\ d_i^y &:= \langle K_x(x^i,\bar{y}) - K_x(x^i,y^i) - K_{xy}(x^i,y^i)(\bar{y}-y^i), x^{i+1}-\bar{x}\rangle \\ &\quad - \langle K_y(x^{i+1},y^{i+1}) - K_y(x^{i+1},y^i) + K_y(\bar{x},\bar{y}) - K_y(\bar{x},y^{i+1}), y^{i+1}-\bar{y}\rangle \\ &\quad - \gamma_{F^*}\|y^{i+1}-\bar{y}\|^2, \end{aligned}$$

we rearrange and estimate

$$\begin{aligned} q_i &= d_i^x + d_i^y + \langle K_y(x^{i+1},y^{i+1}) - K_y(x^{i+1},y^i), y^{i+1}-\bar{y}\rangle \\ &\quad + \langle (K_{xy}(x^{i+1},y^{i+1}) - K_{xy}(x^i,y^i))(y^i-\bar{y}), x^{i+1}-\bar{x}\rangle \\ &\quad + \langle K_x(x^i,y^i) - K_x(x^{i+1},y^{i+1}) + K_{xy}(x^{i+1},y^{i+1})(y^{i+1}-y^i), x^{i+1}-\bar{x}\rangle \\ &\le d_i^x + d_i^y + O(\|u^{i+1}-u^i\|). \end{aligned} \tag{6.5}$$

Using $\xi_x = \gamma_G$, $\xi_y = \gamma_{F^*}$, (3.5), and both Assumption 3.1 and Assumption 3.2 (iv) at $\bar{u}$, we estimate $q_i \le O(\|u^{i+1}-u^i\|)$ as

$$\begin{aligned} d_i^x &\le (\|y^{i+1}-\bar{y}\| - \theta_x)\|K_y(x^{i+1},y^{i+1}) - K_y(\bar{x},y^{i+1}) + K_{yx}(x^{i+1},y^{i+1})(\bar{x}-x^{i+1})\| \le 0, \\ d_i^y &\le (\|x^{i+1}-\bar{x}\| - \theta_y)\|K_x(x^i,\bar{y}) - K_x(x^i,y^i) - K_{xy}(x^i,y^i)(\bar{y}-y^i)\| \le 0. \end{aligned}$$

In the last bounds we used $\theta_x \ge 2\rho_y$, $\theta_y \ge 2\rho_x$, and $\|y^{i+1}-\bar{y}\| \le 2\rho_y$ because both $\|y^{i+1}-\widehat{y}\| \le \rho_y$ and $\|\widehat{y}-\bar{y}\| \le \rho_y$; likewise, $\|x^{i+1}-\bar{x}\| \le 2\rho_x$. Since $\|u^{i+1}-u^i\| \to 0$, we obtain that $\limsup_{i\to\infty} q_i \le 0$. The Brezis–Crandall–Pazy Lemma thus yields the desired inclusion $\bar{v} \in A(\bar{u})$.

Hence in both cases, $\bar{u} \in H^{-1}(0)$ and the condition (ii) of Lemma 6.2 is satisfied. Applying Lemma 6.2, we obtain the claim. □

We now provide convergence rates under additional assumptions of strong convexity of G and/or F^*, although we still allow non-convexity of the overall problem through K. To be specific, we require that we can take the acceleration or steplength update factors $\widetilde{\gamma}_G > 0$ and/or $\widetilde{\gamma}_{F^*} > 0$ in (4.8d) and (4.8e), respectively. Let us start with $\widetilde{\gamma}_G > 0$, which is the case, for instance, when G is strongly convex and (3.4a) holds with $\xi_x = 0$. Since we obtain *a fortiori* strong convergence from the rates, we do not require the additional assumptions on K introduced in Theorem 6.1; on the other hand, we only obtain convergence of the primal iterates. Similar to the linear case of [10], the steplength choice follows directly from having to satisfy (4.8b) and the desire to keep the right-hand side of the σ-rule (4.8c) constant.

Theorem 6.3 (convergence rates under acceleration: $\omega_i = 1$). *Suppose Assumptions 3.1, 3.2 and 5.1 hold for some $R_K > 0$; $L_{yx} \geq 0$; $\lambda_x, \lambda_y, \theta_x, \theta_y \geq 0$; and $\xi_x, \xi_y \in \mathbb{R}$ such that for some $\widetilde{\gamma}_G > 0$,*

$$\xi_x = \gamma_G - \widetilde{\gamma}_G, \qquad \theta_y \geq \rho_x, \tag{6.6a}$$

$$\xi_y = \gamma_{F^*}, \qquad \theta_x \geq \rho_y. \tag{6.6b}$$

Choose

$$\tau_{i+1} = \frac{\tau_i}{1 + 2\widetilde{\gamma}_G \tau_i}, \quad \sigma_{i+1} \equiv \sigma, \quad \text{and} \quad \omega_i \equiv 1, \tag{6.7}$$

satisfying for some $0 < \delta \leq \mu < 1$ the bounds

$$0 < \tau_0 \leq \frac{\delta}{\lambda_x + 3L_{yx}\rho_y} \quad \text{and} \quad 0 < \sigma\tau_0 \leq \frac{1-\mu}{R_K^2}. \tag{6.8}$$

Then $\|x^N - \widehat{x}\|^2$ converges to zero at the rate $O(1/N)$.

Proof. We again first verify (4.8) so that we can apply Theorem 4.2 and Lemma 5.2. Setting $\psi_i \equiv 1$, $\eta_i \equiv \sigma$, $\phi_i := \sigma\tau_i^{-1}$, and $\widetilde{\gamma}_{F^*} = 0$, (4.8a) follows from the σ-rule of (6.7) and the choice of ψ_i, η_i, and ϕ_i. Using (6.7) and $\tau_i := \sigma\phi_i^{-1}$, we obtain $\phi_{i+1} = (1 + 2\widetilde{\gamma}_G\tau_i)\phi_i$, and hence (4.8b) follows. Since $\tau_i \leq \tau_0$ and $\lambda_y \geq 0$, (4.8c) follows from (6.8) and $\omega_i = 1$. Furthermore, (4.8d) and (4.8e) are satisfied due to the assumed bounds (6.6) on ξ_x, ξ_y, θ_x, and θ_y taking $\overline{\omega} = \underline{\omega} = 1$.

We can thus apply Theorem 4.2 and Lemma 5.2 to arrive at (4.7) for $\Delta_{i+1} = 0$. We now estimate the convergence rate from (4.7) by bounding $Z_{N+1}M_{N+1}$ from below. Using Corollary 4.3, we obtain $\delta\phi_N\|x^N - \widehat{x}\|^2 \leq \|u^0 - \widehat{u}\|^2_{Z_1 M_1}$. Moreover,

$$\phi_{N+1} = (1 + 2\widetilde{\gamma}_G\tau_N)\phi_N = \phi_N + 2\widetilde{\gamma}_G\sigma = \ldots = \phi_1 + 2N\widetilde{\gamma}_G\sigma,$$

which yields the claim. □

Theorem 6.4 (linear convergence: $\omega_i < 1$). *Suppose Assumptions 3.1, 3.2 and 5.1 hold for some $R_K > 0$; $L_{yx} \geq 0$; $\lambda_x, \lambda_y \geq 0$; and $\widetilde{\gamma}_G, \widetilde{\gamma}_{F^*} > 0$ as well as*

$$\xi_x = \gamma_G - \widetilde{\gamma}_G, \qquad \theta_y \geq \omega\rho_x, \tag{6.9a}$$

$$\xi_y = \gamma_{F^*} - \widetilde{\gamma}_{F^*}, \qquad \theta_x \geq \rho_y\omega^{-1} \tag{6.9b}$$

with

$$\tau_i \equiv \tau, \quad \sigma_i \equiv \sigma := \tau\widetilde{\gamma}_G\widetilde{\gamma}_{F^*}^{-1}, \quad \text{and} \quad \omega_i \equiv \omega := (1 + 2\widetilde{\gamma}_G\tau)^{-1}. \tag{6.10}$$

Assume for some $0 < \delta \leq \mu < 1$ the bound

$$\tau \leq \min\left\{\frac{\delta}{\lambda_x + 3L_{yx}\rho_y}, \frac{2\widetilde{\gamma}_{F^*}\widetilde{\gamma}_G^{-1}}{\lambda_y + \sqrt{\lambda_y^2 + 4\widetilde{\gamma}_{F^*}\widetilde{\gamma}_G^{-1}(R_K^2(1-\mu)^{-1} + 2\widetilde{\gamma}_G\lambda_y)}}\right\}. \tag{6.11}$$

Then $\|u^N - \widehat{u}\|^2$ converges to zero with the linear rate $O(\omega^N)$.

Proof. We will use Theorem 4.2 and Lemma 5.2, for both of which we need to verify (4.8) first. We set $\overline{\omega} := \underline{\omega} := \omega$,

$$\psi_N := \omega(1 + 2\sigma\widetilde{\gamma}_{F^*})^N = \omega(1 + 2\widetilde{\gamma}_G\tau)^N = \omega^{1-N},$$
$$\phi_N := \omega\sigma\tau^{-1}(1 + 2\tau\widetilde{\gamma}_G)^N = \omega^{1-N}\sigma\tau^{-1}.$$

Then $\psi_1 = 1$ and $\psi_N\sigma = \phi_N\tau$, verifying (4.8a) and (4.8b). We next observe that substituting $\sigma_i = \tau\widetilde{\gamma}_G\widetilde{\gamma}_{F^*}^{-1}$, the first bound of (4.8c) is tantamount to requiring

$$\tau\left(\tau R_K^2(1-\mu)^{-1} + \lambda_y\omega^{-1}\right) \leq \widetilde{\gamma}_{F^*}\widetilde{\gamma}_G^{-1}.$$

Substituting $\omega = (1 + 2\widetilde{\gamma}_G\tau)^{-1}$, this in turn is equivalent to

$$\left(R_K^2(1-\mu)^{-1} + 2\widetilde{\gamma}_G\lambda_y\right)\tau^2 + \lambda_y\tau - \widetilde{\gamma}_{F^*}\widetilde{\gamma}_G^{-1} \leq 0,$$

which after solving a quadratic inequality for τ yields the second bound of (6.11). Since $\omega \leq 1$, the first bound of (6.11) gives the second bound of (4.8c). Finally, (4.8d) and (4.8e) follow directly from (6.9) with $\underline{\omega} = \overline{\omega} = \omega$.

Since Assumption 5.1 and (4.8) hold, we can apply Lemma 5.2 to obtain $\{u^i\}_{i\in\mathbb{N}} \in \mathcal{U}(\rho_x, \rho_y)$ and $\{\bar{x}^{i+1}\}_{i\in\mathbb{N}} \in \mathbb{B}(\widehat{x}, \rho_x)$. Moreover, (4.9) yields self-adjointness of $Z_{i+1}M_{i+1}$. Consequently, we can apply Theorem 4.2 and Lemma 5.2 to arrive at (4.7) for any $\Delta_{i+1} \leq 0$.

We now estimate the convergence rate from (4.7) by bounding $Z_{N+1}M_{N+1}$ from below. Using Corollary 4.3, we obtain that

$$\frac{1}{\omega^N}\left(\delta\sigma\omega\tau^{-1}\|x^N - \widehat{x}\|^2 + \frac{\mu-\delta}{1-\delta}\|y^N - \widehat{y}\|^2\right) \leq \|u^0 - \widehat{u}\|^2_{Z_1M_1}. \tag{6.12}$$

Since $\omega \in (0,1)$, this gives the claimed linear convergence rate through the exponential growth of $1/\omega^N$. □

Remark 6.5. *If $K(x,y) = \langle A(x), y\rangle$ for some $A \in C^1(X)$, then $K_x(x,y) = [\nabla A(x)]^* y$ and $K_y(x,y) = A(x)$ with $L_y(x) = 0$ and $L_{yx} = L$ for L a local Lipschitz factor of ∇A. Furthermore, Assumption 3.2, the steplength bounds, and the update rules required in Theorem 6.1 or 6.4 reduce to the corresponding ones introduced in [10] for this case. As for acceleration, Theorem 6.3 now gives a weaker convergence rate of $O(1/N)$ compared to $O(1/N^2)$ in [10, Theorem 4.3]. This is due to (4.8c) requiring σ_i to be bounded whenever $\lambda_y > 0$, even when τ_i goes to zero.*

Before we conclude this section, we refine Assumption 5.1 by showing that its implicit requirements do not add any additional steplength bounds provided the starting point is sufficiently close to $\widehat{u}$.

Proposition 6.6. *Under the assumptions of Theorem 6.1, 6.3, or 6.4, suppose that $\rho_x, \rho_y > 0$. Then there exists $\varepsilon > 0$ such that Assumption 5.1 holds whenever the initial iterate $u^0 = (x^0, y^0)$ satisfies*

$$r_{\max} := \sqrt{2\delta^{-1}(\|x^0 - \widehat{x}\|^2 + \nu^{-1}\|y^0 - \widehat{y}\|^2)} \leq \varepsilon \quad \text{with} \quad \nu := \sigma_1\omega_0\tau_0^{-1}. \tag{6.13}$$

Proof. We take μ, δ, σ_i, τ_i, and ω_i as they are defined in the corresponding Theorem 6.1, 6.3, or 6.4, and $L_x(\widehat{y})$, $L_y(\widehat{x})$, R_K from Assumption 3.2. We need to show that there exist $\delta_x, \delta_y > 0$ and $r_y \geq r_{\max}\sqrt{\nu(1-\delta)\delta(\mu-\delta)^{-1}}$ such that (5.7) holds and

$$\mathbb{B}(\widehat{x}, r_{\max} + \delta_x) \times \mathbb{B}(\widehat{y}, r_y + \delta_y) \subseteq \mathcal{U}(\rho_x, \rho_y). \tag{6.14}$$

Let $\varepsilon > 0$ and set $r_y := \varepsilon\sqrt{\nu(1-\delta)\delta(\mu-\delta)^{-1}}$ as well as $\delta_x := \sqrt{\varepsilon}$ and $\delta_y := \rho_y - r_y$. Observing (6.13), we then see both that $\delta_y > 0$ and that (6.14) holds for $\varepsilon > 0$ sufficiently small. Furthermore, (6.13) yields that $r_{\max} \leq \varepsilon$ in Lemma 5.2. Let

$$c_\varepsilon := \min\left\{\frac{\delta_x}{2R_K r_y + 2L_x(\widehat{y})r_{\max}}, \frac{\delta_y}{L_y(\widehat{x})r_y + R_K(r_{\max} + \delta_x)}\right\}.$$

Since $r_y, r_{\max} = O(\varepsilon)$, $\delta_x = \sqrt{\varepsilon}$, and $\delta_y > \rho_y/2 > 0$ for $\varepsilon > 0$ small enough, we see that $c_\varepsilon \to \infty$ as $\varepsilon \to 0$. Comparing the definition of c_ε to (5.7), we therefore see that the latter holds for any given $\tau_0 > 0$ and $\sigma_i \equiv \sigma > 0$ by taking $\varepsilon > 0$ sufficiently small. Since in Theorems 6.1, 6.3 and 6.4 we have $\tau_i \leq \tau_0$, the inequalities (5.7) hold. □

7 NUMERICAL EXAMPLES

Finally, we illustrate the applicability of the proposed approach for the example applications described in Section 2. The Julia implementation used to generate the following results is on Zenodo [11].

7.1 AN ELLIPTIC NASH EQUILIBRIUM PROBLEM

Our first example illustrates the reformulation from Section 2.1 for the two-player elliptic Nash equilibrium problem from [6]. Here the action space of each player is $L^2(\Omega)$ for a bounded domain $\Omega \subset \mathbb{R}^d$ with boundary $\partial\Omega$. To avoid confusion with the spatial variable, we will in this subsection denote the primal variable with u and the dual variable with v. The set of admissible strategies is

$$X_k = \left\{w \in L^2(\Omega) : w(x) \in [a, b] \text{ a.e. } x \in \Omega\right\} \qquad (k = 1, 2).$$

For a set of strategies $u := (u_1, u_2) \in X = X_1 \times X_2$, the payout function for each player is

$$\phi_k(u_1, u_2) = \frac{1}{2}\|S(u_1, u_2) - z_k\|^2_{L^2(\Omega)} + \frac{\alpha_k}{2}\|B_k u_k\|^2_{L^2(\Omega)} \qquad (k = 1, 2),$$

where $\alpha_k > 0$, $z_k \in L^2(\Omega)$ are given target states, $S : L^2(\Omega)^2 \to L^2(\Omega)$ maps $u = (u_1, u_2)$ to the solution y to the elliptic boundary value problem

$$\begin{cases} -\Delta y = B_1 u_1 + B_2 u_2 + f & \text{on } \Omega, \\ y = 0 & \text{on } \partial\Omega, \end{cases} \tag{7.1}$$

$B_k : L^2(\Omega) \to L^2(\Omega)$ are control operators which are here chosen as

$$B_k w := \begin{cases} w(x) & \text{if } x \in \omega_k, \\ 0 & \text{if } x \notin \omega_k, \end{cases}$$

for some control domains $\omega_k \subset \Omega$, and f is a common source term. Following Section 2.1, the corresponding Nash equilibrium problem (2.1) can then be solved by applying Algorithm 1.1 to

$$\begin{aligned} G &: L^2(\Omega)^2 \to \overline{\mathbb{R}}, & G(u_1, u_2) &= \delta_X(u_1, u_2), \\ F^* &: L^2(\Omega)^2 \to \overline{\mathbb{R}}, & F^*(v_1, v_2) &= \delta_X(v_1, v_2), \\ K &: L^2(\Omega)^2 \times L^2(\Omega)^2 \to \mathbb{R}, & K((u_1, u_2), (v_1, v_2)) &= [\phi_1(u_1, u_2) - \phi_1(v_1, u_2)] \\ & & &\quad + [\phi_2(u_1, u_2) - \phi_2(u_1, v_2)] \end{aligned}$$

To implement the algorithm, we need explicit forms of the proximal mappings for G and F^* and of the partial derivatives of K. Since $G = F^* = \delta_X$ for $X = X_1 \times X_2$, we have

$$\operatorname{prox}_{\tau G}(w) = \operatorname{prox}_{\sigma F^*}(w) = \operatorname{proj}_X(w) = \Big(\operatorname{proj}_{X_1}(w_1), \operatorname{proj}_{X_2}(w_2)\Big)$$

for the metric projections onto the convex sets X_k given pointwise almost everywhere by

$$[\operatorname{proj}_{X_k}(w_k)](x) = \begin{cases} b & \text{if } w_k(x) > b, \\ w_k(x) & \text{if } w_k(x) \in [a, b], \\ a & \text{if } w_k(x) < a. \end{cases}$$

It remains to address the computation of $K_u(u, v)$ and $K_v(u, v)$. Using adjoint calculus and the linearity of the adjoint equation, we have that

$$K_u(u, v) = \begin{pmatrix} p_1(u, v) + \alpha_1 u_1 \\ p_2(u, v) + \alpha_2 u_2 \end{pmatrix}, \qquad K_v(u, v) = \begin{pmatrix} q_1(u, v) - \alpha_1 v_1 \\ q_2(u, v) - \alpha_2 v_2 \end{pmatrix},$$

where $p_1(u, v) =: p_1$ and $p_2(u, v) =: p_2$ are the solutions to the equations

$$\begin{aligned} -\Delta p_1 &= 2S(u_1, u_2) - S(u_1, v_2) - z_1, \\ -\Delta p_2 &= 2S(u_1, u_2) - S(v_1, u_2) - z_2, \end{aligned}$$

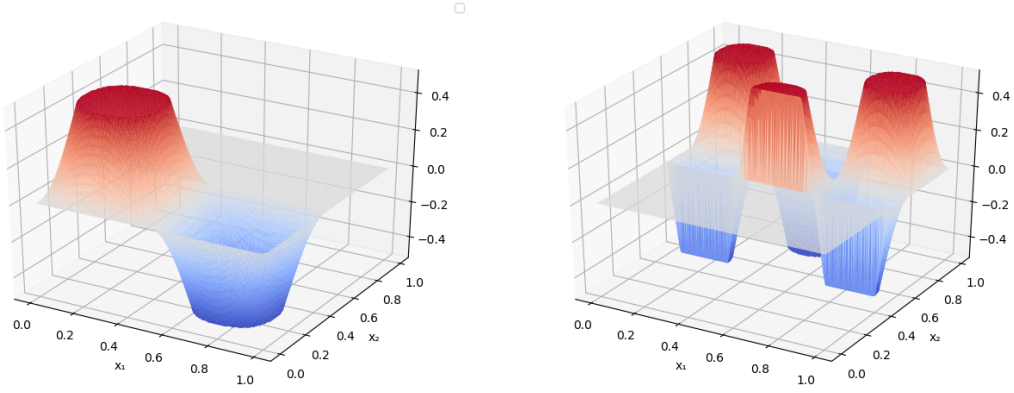
and $q_1(u, v) =: q_1$ and $q_2(u, v) =: q_2$ are the solutions to the equations

$$\begin{aligned} -\Delta q_1 &= -S(v_1, u_2) + z_1, \\ -\Delta q_2 &= -S(u_1, v_2) + z_2, \end{aligned}$$

all with homogeneous Dirichlet conditions. Hence, every iteration of Algorithm 1.1 requires nine solutions of a partial differential equation (recall that K_v is evaluated at $(\bar{u}^{i+1}, v^i)$, while K_u is evaluated at (u^i, v^i)). Since S and hence K_u and K_v are affine in u and v, the assumptions of Theorem 6.1 are satisfied for sufficiently small step sizes. Since neither F^* nor G are strongly convex, no acceleration is possible.

For our numerical tests we follow [6] and consider a finite-difference discretization of (7.1) on $\Omega = (0, 1)^2$ with N nodes in each direction,

$$\omega_1 = (0, 1) \times (0, 1/2), \qquad \omega_2 = (0, 1) \times (1/2, 1),$$



(a) u_1^* (b) u_2^*

Figure 3: Constructed solution for elliptic NEP example ($N = 128$)

Table 1: Results for elliptic NEP example for different N

i	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
1	$1.298 \cdot 10^{-1}$	$1.319 \cdot 10^{-1}$	$1.330 \cdot 10^{-1}$	$1.335 \cdot 10^{-1}$	$1.338 \cdot 10^{-1}$
2	$3.889 \cdot 10^{-6}$	$4.048 \cdot 10^{-6}$	$4.074 \cdot 10^{-6}$	$4.088 \cdot 10^{-6}$	$4.097 \cdot 10^{-6}$
3	$3.835 \cdot 10^{-10}$	$3.977 \cdot 10^{-10}$	$4.010 \cdot 10^{-10}$	$4.026 \cdot 10^{-10}$	$4.032 \cdot 10^{-10}$
4	$3.811 \cdot 10^{-14}$	$3.952 \cdot 10^{-14}$	$3.986 \cdot 10^{-14}$	$4.001 \cdot 10^{-14}$	$4.008 \cdot 10^{-14}$
5	$3.787 \cdot 10^{-18}$	$3.928 \cdot 10^{-18}$	$3.963 \cdot 10^{-18}$	$3.977 \cdot 10^{-18}$	$3.985 \cdot 10^{-18}$

as well as $a = -0.5$, $b = 0.5$, and $\alpha_i = 1$. Using the method of manufactured solutions, z_1, z_2, and f are chosen such that the solution $u^* = (u_1^*, u_2^*)$ of the Nash equilibrium problem is known a priori; see Figure 3. By construction, the saddle point then satisfies $v^* = u^*$ and hence $\Psi(u^*, v^*) = 0$.

Since the Lipschitz constants for K and its derivatives are not available, we simply take the parameters in Algorithm 1.1 as $\sigma_{i+1} \equiv \sigma = 1.0$, $\tau_i \equiv \tau = 0.99$, and $\omega = 1.0$. We initialize with $u^0 = v^0 \equiv 0$. Although these choices do not satisfy the requirements of our convergence analysis, for this simple model problem we nevertheless observe that the algorithm performs well. The results of the algorithm for different values of $N \in \{64, 128, 256, 512, 1024\}$ are shown in Table 1, which reports the distance of the primal-dual iterates (u^i, v^i) to the exact solution. As can be seen, the iteration converges in each case to machine precision within 5 iterations, and the convergence behavior is virtually identical. This demonstrates the mesh independence expected from an algorithm for which convergence can be shown in function spaces.

7.2 ℓ^0-TV DENOISING

Our next example concerns the ℓ^0-TV denoising or segmentation problem from Section 2.2. Recall that we can solve the (Huber-regularized) ℓ^0-TV problem (2.3) by applying Algorithm 1.1

to

$$
\begin{aligned}
G &: \mathbb{R}^{N_1\times N_2} \to \mathbb{R}, & G(x) &= \frac{1}{2\alpha}\|x-f\|_2^2,\\
F_\gamma^* &: \mathbb{R}^{N_1\times N_2\times 2} \to \mathbb{R}, & F_\gamma^*(y) &= \frac{\gamma}{2}\|y\|_2^2,\\
K_p &: \mathbb{R}^{N_1\times N_2} \times \mathbb{R}^{N_1\times N_2\times 2} \to \mathbb{R}, & K_p(x,y) &= \kappa_p(D_h x, y),
\end{aligned}
$$

for $p \in \{1,\infty\}$ and $\gamma \geq 0$, where $D_h : \mathbb{R}^{N_1\times N_2} \to \mathbb{R}^{N_1\times N_2\times 2}$ is the discrete gradient. We write H_γ for H defined in (4.1) corresponding to $F^* = F_\gamma^*$. Since G and F_γ^* are quadratic, a simple computation shows that

$$
\operatorname{prox}_{\tau G}(x) = \frac{1}{1+\frac{\tau}{\alpha}}\left(x + \frac{\tau}{\alpha} f\right), \quad \text{and} \quad \operatorname{prox}_{\sigma F_\gamma^*}(y) = \frac{1}{1+\gamma\sigma} y,
$$

where all operations are to be understood componentwise. For the derivatives of K_p, we have by the chain rule

$$
K_x(x,y) = D_h^T \kappa_{p,z}(D_h x, y), \qquad K_y(x,y) = \kappa_{p,y}(D_h x, y), \tag{7.2}
$$

where D_h^T is the discrete (negative) divergence. For the partial derivatives of $\kappa_{p,z}(z,y)$ and $\kappa_{p,y}(z,y)$, we again distinguish the cases $p = 1$ and $p = \infty$:

For $p = 1$, we have componentwise

$$
\begin{aligned}
[\kappa_{1,z}(z,y)]_{ijk} &= 2(1 - z_{ijk}y_{ijk})y_{ijk},\\
[\kappa_{1,y}(z,y)]_{ijk} &= 2(1 - z_{ijk}y_{ijk})z_{ijk}.
\end{aligned}
$$

For $p = \infty$, we have componentwise

$$
\begin{aligned}
[\kappa_{\infty,z}(z,y)]_{ijk} &= 2(1 - z_{ij1}y_{ij1} - z_{ij2}y_{ij2})y_{ijk},\\
[\kappa_{\infty,y}(z,y)]_{ijk} &= 2(1 - z_{ij1}y_{ij1} - z_{ij2}y_{ij2})z_{ijk}.
\end{aligned}
$$

It remains to choose valid step sizes for Algorithm 1.1, for which the next result gives useful estimates. We recall from [7] that a forward differences discretization of the gradient operator satisfies $\|D_h\|_2 \leq \sqrt{8}/h$. Recalling (7.2) and the definitions of G and F_γ^*, a critical point $(\widehat{x},\widehat{y}) \in H_\gamma^{-1}(0)$ satisfies

$$
0 = \alpha^{-1}(\widehat{x} - f) + D_h^T \kappa_{p,z}(D_h\widehat{x}, \widehat{y}) \quad \text{and} \quad \gamma\widehat{y} = \kappa_{p,y}(D_h\widehat{x}, \widehat{y}). \tag{7.3}
$$

For brevity, we set

$$
\begin{aligned}
\widehat{m}_x &:= \max_{ij} |[D_h\widehat{x}]_{\cdot ij}|_2 \quad \text{and} & \widehat{m}_y &:= \max_{ij} |\widehat{y}_{\cdot ij}|_2 & (p=\infty),\\
\widehat{m}_x &:= \max_{kij} |[D_h\widehat{x}]_{kij}| \quad \text{and} & \widehat{m}_y &:= \max_{kij} |\widehat{y}_{kij}| & (p=1).
\end{aligned}
$$

Using the results of Appendix C we verify the fundamental Assumption 3.2.

Corollary 7.1. *Let $K = K_p$ for either $p = 1$ or $p = \infty$. Choose $L \geq \|D_h\|_2$ and $R_K > 2L$. Then Assumption 3.2 holds for some $\theta_x, \theta_y > 0$ and $\rho_x, \rho_y > 0$ with*

$$L_x(y) = 2L^2\|y\|_2^2, \qquad L_y(x) = 2L^2\|x\|_2^2, \qquad L_{yx} = 4L \sup_{y \in \mathbb{B}(\widehat{y}, \rho_y)} \|y\|_2,$$

and the constants $\xi_x, \xi_y > 0$, $\lambda_x, \lambda_y \geq 0$ satisfying

$$\xi_x \lambda_x > 2L^2(L^{-1}\lambda_x + \widehat{m}_y^2)\widehat{m}_y^2 \quad \text{and} \quad \lambda_y > \widehat{m}_x^2. \tag{7.4}$$

Proof. We consider only $p = \infty$ as the proof for $p = 1$ is similar. Taking $\widetilde{R}_K > 2$, Lemma C.1 applied componentwise shows that the operator κ_p satisfies Assumption 3.2 for some $\widetilde{\theta}_z, \widetilde{\theta}_y > 0$ and $\widetilde{\rho}_x, \widetilde{\rho}_y > 0$ (depending on $\widetilde{R}_K$) when we take

$$\widetilde{L}_z(y) = 2\|y\|_2^2, \qquad \widetilde{L}_y(z) = 2\|z\|_2^2, \quad \text{and} \quad \widetilde{L}_{yz} = 4 \max_{y \in \mathbb{B}(\widehat{y}, \widetilde{\rho_y})} \|y\|_2.$$

Moreover, the constants $\widetilde{\xi}_z, \widetilde{\xi}_y \in \mathbb{R}$ and $\widetilde{\lambda}_z, \widetilde{\lambda}_y \geq 0$ need to satisfy $\widetilde{\xi}_z \widetilde{\lambda}_z > \max_{ij} 2(\lambda_z + \|\widehat{y}_{\cdot ij}\|^2)\|\widehat{y}_{\cdot ij}\|^2$ as well as $\widetilde{\xi}_y > 0$ and $\widetilde{\lambda}_y > \max_{ij} \|\widehat{z}_{\cdot ij}\|^2$ for $\widehat{z} = D_h\widehat{x}$.

By Lemma C.2 on compositions with a linear operator, we can now take

$$\begin{aligned}
&R_K = \widetilde{R}_K L, && \rho_x = L^{-1}\widetilde{\rho}_x, && \rho_y = \widetilde{\rho}_y, && \xi_x = L\widetilde{\xi}_z, && \xi_y = \widetilde{\xi}_y,\\
&\lambda_x = L\widetilde{\lambda}_z, && \lambda_y = \widetilde{\lambda}_y, && \theta_x = \widetilde{\theta}_z, && \theta_y = \widetilde{\theta}_y L^{-1}, &&\\
&L_x(y) = L^2\widetilde{L}_z(y), && L_y(x) = \widetilde{L}_y(D_h x), && L_{yx} = L^2\widetilde{L}_{yz}. &&
\end{aligned}$$

These give the claim. □

We now obtain from Theorem 6.4 the following estimate.

Corollary 7.2. *Suppose Assumption 3.1 holds. Choose $L \geq \|D_h\|_2$. For some $\widetilde{\gamma}_G \in (0, \alpha^{-1})$ and $\widetilde{\gamma}_{F^*} \in (0, \gamma)$, take $\xi_x = \alpha^{-1} - \widetilde{\gamma}_G$ and $\xi_y = \gamma - \widetilde{\gamma}_{F^*}$ as well as $\lambda_x, \lambda_y \geq 0$ such that (7.4) holds. For some $0 < \delta \leq \mu < 1$, take $\sigma = \tau\widetilde{\gamma}_G\widetilde{\gamma}_{F^*}^{-1}$ and $\omega := (1 + 2\widetilde{\gamma}_G\tau)^{-1}$ as well as*

$$\tau < \min\left\{\frac{\delta}{\lambda_x}, \frac{2\widetilde{\gamma}_{F^*}\widetilde{\gamma}_G^{-1}}{\lambda_y + \sqrt{\lambda_y^2 + 4\widetilde{\gamma}_{F^*}\widetilde{\gamma}_G^{-1}(4L^2(1-\mu)^{-1} + 2\widetilde{\gamma}_G\lambda_y)}}\right\}. \tag{7.5}$$

Then $\|u^N - \widehat{u}\|^2$ converges to zero with the linear rate $O(\omega^N)$ provided u^0 is close enough to $\widehat{u}$.

Proof. The assumptions $\widetilde{\gamma}_G \in (0, \alpha^{-1})$ and $\widetilde{\gamma}_{F^*} \in (0, \gamma)$ ensure $\xi_x, \xi_y > 0$. Since we have assumed (7.4), Corollary 7.1 yields Assumption 3.2 for any $R_K > 2L$ and some $\theta_x, \theta_y > 0$. We next use Theorem 6.4, whose conditions we need to verify. First, taking $\rho_x, \rho_y > 0$ ensures that $\theta_x \geq \rho_y\omega^{-1}$ and $\theta_y \geq \omega\rho_x$. Furthermore, the strict inequality in (7.5) implies (6.11) for sufficiently small $\rho_y > 0$. Finally, Proposition 6.6 ensures that we can satisfy Assumption 5.1 by taking u^0 sufficiently close to $\widehat{u}$. The rest of the conditions we have assumed explicitly, so we can apply Theorem 6.4 to finish the proof. □

Recall that Assumption 3.1 is a second-order growth condition at the critical point $(\widehat{x}, \widehat{y})$, which is a common assumption needed to show convergence of algorithms for non-convex optimization problems. To calculate the upper bounds on τ in (7.5), we need to find $\lambda_x, \lambda_y \geq 0$ satisfying (7.4). For this, in turn, we need to estimate $\widehat{m}_x$ and $\widehat{m}_y$. To do this, note that the critical point conditions (7.3) imply

$$\widehat{y}_{\cdot ij} = \frac{2[D_h \widehat{x}]_{\cdot ij}}{2|[D_h \widehat{x}]_{\cdot ij}|_2^2 + \gamma} \quad (p = \infty) \quad \text{and} \quad \widehat{y}_{kij} = \frac{2[D_h \widehat{x}]_{kij}}{2|[D_h \widehat{x}]_{kij}|^2 + \gamma} \quad (p = 1). \tag{7.6}$$

Since $t \mapsto t/(t+\gamma)$ is increasing, we can estimate $\widehat{m}_y$ based on $\widehat{m}_x$. Since any solution of the Potts problem should be piecewise constant with very few intensity quantization levels, we can estimate $\widehat{m}_x$ as the expected maximal jump between neighboring pixels. We take this as 100% of the dynamic range for safety. In practice, as a practical choice of $\gamma > 0$ will likely not satisfy $\xi_x > 2L\widehat{m}_y^2$, we use an over-approximation $\overline{\gamma} := 10 \geq \gamma$ in (7.6). We remark that we thus cannot guarantee convergence of Algorithm 1.1 for small $\gamma > 0$; however, we demonstrate below that these estimates can still lead to useful step sizes for such cases. Similarly, we do not have an estimate for the unknown local neighborhood of convergence; we compensate for this by taking small $\delta = 0.1$ in (7.5). As the results below demonstrate, with these parameters we nevertheless observe convergence for the reasonable starting point $u^0 = (x^0, y^0)$ with $x^0 = f$ and $y^0 \equiv 0$.

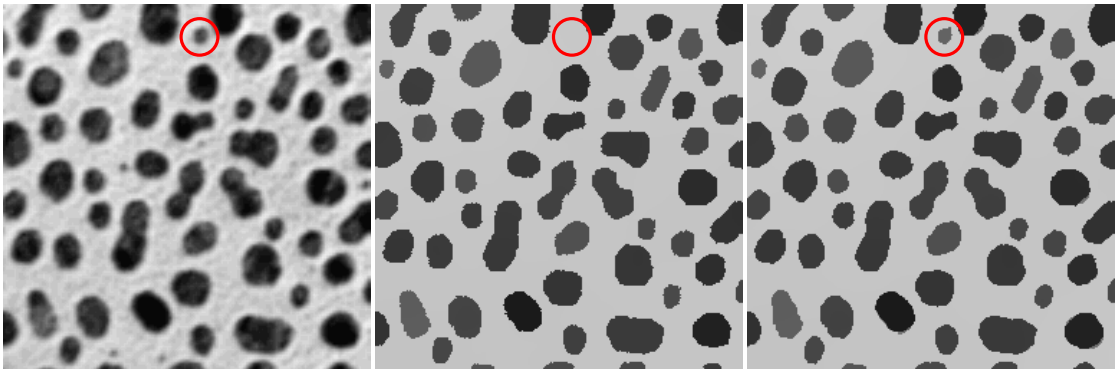
We illustrate the performance of the algorithm and the effects of the choice of p. As a test image, we choose "blobs" from the ImageJ framework [30] with size $N_1 \times N_2 = 256 \times 254$, see Figure 4a. We set $\alpha = 1$ and $\gamma = 10^{-3}$ (cf. Figure 2) and use the accelerated step size rule from Theorem 6.4. To do this, we need to satisfy (7.5) for the primal steplength τ. We discretize the problem such that $h = 1$ and hence $L = \sqrt{8}$. Furthermore, we set $\widetilde{\gamma}_{F^*} = \gamma/100$ and $\widetilde{\gamma}_G = \widetilde{\alpha}^{-1}$ for $\widetilde{\alpha} = 10\alpha$. The above estimates then lead to the steplength parameters

$p = 1$: $\tau = 1.04085 \cdot 10^{-3}$, $\sigma = 1.04085$, $\omega = 0.99480$;

$p = \infty$: $\tau = 5.51922 \cdot 10^{-4}$, $\sigma = 0.551922$, $\omega = 0.99724$.

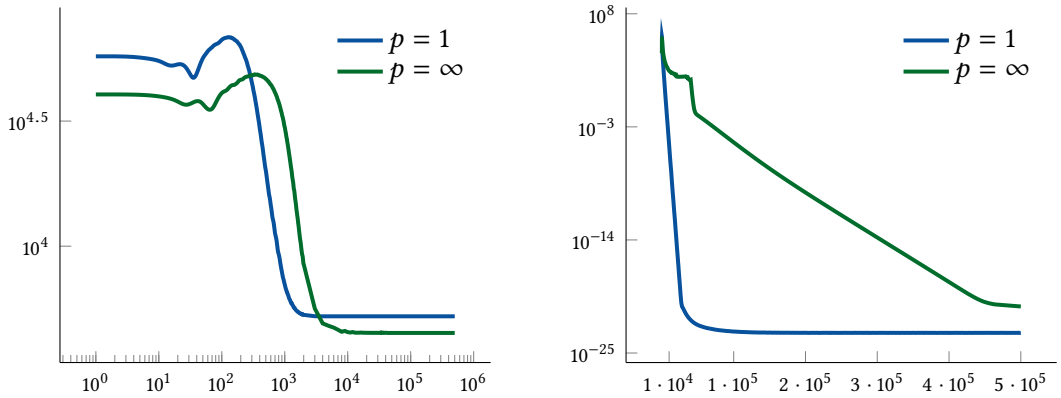
Since the exact solution $(\widehat{x}, \widehat{y})$ is not available here, we instead use $x^{\max} := x^{N_{\max}}$ for $N_{\max} = 10^6$ and similarly $y^{\max}$ as references for computing errors. The corresponding reference images $x^{\max}$ obtained from Algorithm 1.1 after $N_{\max} = 10^6$ iterations are shown in Figures 4b and 4c for $p = 1$ and $p = \infty$, respectively. While the evaluation of the formulation and the algorithm in the context of image processing is outside of the scope of this work, we briefly comment on the difference between $p = 1$ and $p = \infty$. As can be seen by comparing the two images, the results are very similar. However, since diagonal jumps are penalized less for $p = \infty$, the isotropic Huber–Potts model is better able to preserve small light blobs such as the one indicated by the red circles. The edges of the blobs are also noticeably smoother.

The convergence behavior of the method for both choices of p over $N_{\max}/2 = 5 \cdot 10^5$ iterations is given in Figure 5. For the function values, we observe in Figure 5a the usual fast decrease in the beginning of the iteration, after which the values stagnate. Nevertheless, the errors continue to decrease down to machine precision at the predicted linear rate. The convergence behavior for $p = 1$ and $p = \infty$ is similar, although the linear convergence for $p = \infty$ is with a significantly smaller constant. We remark that visually, the iterates in both cases are indistinguishable from



(a) original image f (b) $x^{\max}$ for $p = 1$ (c) $x^{\max}$ for $p = \infty$

Figure 4: ℓ^0-TV denoising: original image f and reference iterates $x^{\max}$ for anisotropic ($p = 1$) and isotropic ($p = \infty$) Huber–Potts model



(a) function values (b) errors

Figure 5: ℓ^0-TV denoising: convergence of function values $F_{\gamma}(x^N) + G(x^N)$ and errors $\|x^N - x^{\max}\|^2 + \|y^N - y^{\max}\|^2$

the reference images already after $N = 10^4$ iterations. This is consistent with Figure 5b since the total error is dominated by the dual component, which acts as an edge indicator; small changes of the boundaries of the blobs during the iteration will, even for small gray value changes, lead to large differences in the dual variable.

8 CONCLUSION

Using generalized conjugation, some non-smooth non-convex optimization problems can be transformed into saddle-point problems involving non-smooth convex functionals and a smooth non-convex-concave coupling term. For such problems, a generalized primal–dual proximal splitting method can be applied that converges weakly under steplength conditions if a local quadratic growth condition is satisfied near a saddle-point. Under additional strong convexity

assumptions on the functionals (but not the coupling term and hence the problem), convergence rates for accelerated algorithms can be shown. This approach can be applied to elliptic Nash equilibrium problems and for the anisotropic and isotropic Huber-regularized Potts models, as the numerical examples illustrate. Future work is concerned with further evaluating and comparing the performance of the proposed algorithm for these examples.

ACKNOWLEDGMENTS

In the first stages of the research T. Valkonen and S. Mazurenko were supported by the EPSRC First Grant EP/P021298/1, "PARTIAL Analysis of Relations in Tasks of Inversion for Algorithmic Leverage". Later T. Valkonen was supported by the Academy of Finland grants 314701 and 320022. C. Clason was supported by the German Science Foundation (DFG) under grant Cl 487/2-1. We thank the anonymous reviewers for insightful comments.

A DATA STATEMENT FOR THE EPSRC

The source codes for the numerical experiments are on Zenodo at [11].

APPENDIX A REDUCTIONS OF THE THREE-POINT CONDITION

The following two propositions demonstrate that Assumption 3.2 (iv) is closely related to standard second-order optimality conditions, i.e., that the Hessian is positive definite at the solution $\widehat{u}$.

Proposition A.1. *Suppose Assumption 3.2 (ii) (locally Lipschitz gradients of K) holds in some neighborhood $\mathcal{U}$ of $\widehat{u}$, and for some $\xi_x \in \mathbb{R}$, $\gamma_x > 0$,*

$$\xi_x \|x - \widehat{x}\|^2 + \langle K_x(x, \widehat{y}) - K_x(\widehat{x}, \widehat{y}), x - \widehat{x}\rangle \geq \gamma_x \|x - \widehat{x}\|^2 \quad ((x, y) \in \mathcal{U}). \tag{A.1}$$

Then (3.4a) holds in $\mathcal{U}$ with $\theta_x = 2(\gamma_x - \alpha)L_{yx}^{-1}$, and $\lambda_x = L_x(\widehat{y})^2(2\alpha)^{-1}$ for any $\alpha \in (0, \gamma_x]$.

Proof. An application of Cauchy's and Young's inequalities with any factor $\alpha > 0$, Assumption 3.2 (ii), and (A.1) yields the estimate

$$\begin{aligned}\langle K_x(x', \widehat{y}) - K_x(\widehat{x}, \widehat{y}), x - \widehat{x}\rangle + \xi_x \|x - \widehat{x}\|^2 &= \langle K_x(x, \widehat{y}) - K_x(\widehat{x}, \widehat{y}), x - \widehat{x}\rangle + \xi_x \|x - \widehat{x}\|^2 \\ &\quad + \langle K_x(x', \widehat{y}) - K_x(x, \widehat{y}), x - \widehat{x}\rangle \\ &\geq (\gamma_x - \alpha)\|x - \widehat{x}\|^2 - L_x(\widehat{y})^2(4\alpha)^{-1}\|x' - x\|^2.\end{aligned}$$

At the same time, using (3.5),

$$\|K_y(\widehat{x}, y) - K_y(x, y) - K_{yx}(x, y)(\widehat{x} - x)\| \leq \frac{L_{yx}}{2}\|x - \widehat{x}\|^2.$$

Therefore (3.4a) holds if we take $\theta_x \leq 2(\gamma_x - \alpha)L_{yx}^{-1}$ and $\lambda_x = L_x(\widehat{y})^2(2\alpha)^{-1}$. □

Proposition A.2. *Suppose Assumption 3.2 (ii) (locally Lipschitz gradients of K) holds in some neighborhood $\mathcal{U}$ of $\widehat{u}$ with $L_y(x) \le \bar{L}_y$, and that*

$$\|K_{xy}(x,y') - K_{xy}(x,y)\| \le L_{xy}\|y' - y\| \quad (u, u' \in \mathcal{U})$$

for some constant $L_{xy} \ge 0$. Assume, moreover, for some $\xi_y \in \mathbb{R}$, $\gamma_y > 0$ that

$$\xi_y\|y - \widehat{y}\|^2 + \langle K_y(\widehat{x},\widehat{y}) - K_y(\widehat{x},y), y - \widehat{y}\rangle \ge \gamma_y\|y - \widehat{y}\|^2 \quad ((x,y) \in \mathcal{U}). \tag{A.2}$$

Then (3.4b) holds in $\mathcal{U}$ with $\theta_y = 2(\gamma_y - \alpha_1)(1+\alpha_2)^{-1}L_{xy}^{-1}$, and $\lambda_y = (\bar{L}_y^2(2\alpha_1)^{-1} + (1+\alpha_2^{-1})L_{xy}\theta_y)$ for any $\alpha_1 \in (0, \gamma_y]$, $\alpha_2 > 0$.

Proof. An application of Cauchy's and Young's inequalities with any factor $\alpha > 0$, Assumption 3.2 (ii), and (A.2) yields the estimate

$$\begin{aligned}\langle K_y(x,y) - K_y(x,y') + K_y(\widehat{x},\widehat{y}) - K_y(\widehat{x},y), y - \widehat{y}\rangle + \xi_y\|y - \widehat{y}\|^2 \\ \ge \langle K_y(x,y) - K_y(x,y'), y - \widehat{y}\rangle + \gamma_y\|y - \widehat{y}\|^2 \\ \ge (\gamma_y - \alpha_1)\|y - \widehat{y}\|^2 - \frac{L_y(x)^2}{4\alpha_1}\|y' - y\|^2.\end{aligned}$$

At the same time, using (3.5) and Young's inequality for any $\alpha_2 > 0$,

$$\begin{aligned}\|K_x(x',\widehat{y}) - K_x(x',y') - K_{xy}(x',y')(\widehat{y} - y')\| &\le \frac{L_{xy}}{2}\|y' - \widehat{y}\|^2 \\ &\le \frac{L_{xy}}{2}(1+\alpha_2)\|y - \widehat{y}\|^2 + \frac{L_{xy}}{2}(1+\alpha_2^{-1})\|y' - y\|^2.\end{aligned}$$

Therefore (3.4b) holds if we take $\theta_y \le 2\frac{\gamma_y - \alpha_1}{(1+\alpha_2)L_{xy}}$ and $\lambda_y = \frac{\bar{L}_y^2}{2\alpha_1} + (1+\alpha_2^{-1})L_{xy}\theta_y$. □

APPENDIX B RELAXATIONS OF THE THREE-POINT CONDITION

In all the results of this paper, Assumption 3.2 (iv) can be generalized to the following three-point condition similar to the one used in [10].

Assumption B.1. The functional $K(x,y) \in C^1(X \times Y)$ and there exists a neighborhood

$$\mathcal{U}(\rho_x, \rho_y) := (\mathbb{B}(\widehat{x}, \rho_x) \cap \mathcal{X}_G) \times (\mathbb{B}(\widehat{y}, \rho_y) \cap \mathcal{Y}_{F^*}), \tag{B.1}$$

for some $\rho_x, \rho_y > 0$ such that for all $u', u \in \mathcal{U}(\rho_x, \rho_y)$, the following property holds:

(iv*) (three-point condition) There exist $\theta_x, \theta_y > 0$, $\lambda_x, \lambda_y \ge 0$, $\xi_x, \xi_y \in \mathbb{R}$, and $p_x, p_y \in [1,2]$ such that

$$\begin{aligned}&\langle K_x(x',\widehat{y}) - K_x(\widehat{x},\widehat{y}), x - \widehat{x}\rangle + \xi_x\|x - \widehat{x}\|^2 \\ &\quad \ge \theta_x\|K_y(\widehat{x},y) - K_y(x,y) - K_{yx}(x,y)(\widehat{x} - x)\|^{p_x} - \frac{\lambda_x}{2}\|x - x'\|^2,\end{aligned} \tag{B.2A}$$

$$\begin{aligned}&\langle K_y(x,y) - K_y(x,y') + K_y(\widehat{x},\widehat{y}) - K_y(\widehat{x},y), y - \widehat{y}\rangle + \xi_y\|y - \widehat{y}\|^2 \\ &\quad \ge \theta_y\|K_x(x',\widehat{y}) - K_x(x',y') - K_{xy}(x',y')(\widehat{y} - y')\|^{p_y} - \frac{\lambda_y}{2}\|y - y'\|^2.\end{aligned} \tag{B.2B}$$

This assumption introduces p_x and p_y in $[1,2]$, while in Assumption 3.2 (iv) we had $p_x = p_y = 1$. For instance, in [10, Appendix B] we verified Assumption B.1 with $p_x = 2$ for the case $K(x,y) = \langle K(x), y\rangle$ for the reconstruction of the phase and amplitude of a complex number. This relaxation mainly affects the proof of Step 4 in Theorem 4.2, which now requires a few intermediate derivations.

Corollary B.1. *The results of Theorem 4.2 continue to hold if Assumption 3.2 (iv) is replaced with Assumption B.1 (iv*) for some $p_x, p_y \in [1,2]$, where in case $p_y \in (1,2]$, (4.8d) is replaced by*

$$\gamma_G \geq \widetilde{\gamma}_G + \xi_x + \frac{p_y - 1}{(\theta_y p_y^{p_y} \rho_x^{p_y-2} \overline{\omega}^{-1})^{\frac{1}{p_y-1}}}, \tag{B.3A}$$

and in case $p_x \in (1,2]$, (4.8e) is replaced by

$$\gamma_{F^*} \geq \widetilde{\gamma}_{F^*} + \xi_y + \frac{p_x - 1}{(\underline{\omega}\theta_x p_x^{p_x} \rho_y^{p_x-2})^{\frac{1}{p_x-1}}}. \tag{B.3B}$$

Proof. The beginning of the proof follows the exact same steps as in the proof of Theorem 4.2 up until (4.14). We now use Assumption B.1 (iv*) to further bound D_x and D_y similarly to (4.15) and (4.16). From (B.2A),

$$\begin{aligned} D_x \geq \theta_x \|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\|^{p_x} - \frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 \\ - \|y^{i+1} - \widehat{y}\|\|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\|\omega_i^{-1}. \end{aligned} \tag{B.4}$$

The following generalized Young's inequality for any positive a, b, p and q such that $q^{-1} + p^{-1} = 1$ allows for our choice of varying $p_x \in [1,2]$:

$$ab = \left(ab^{\frac{2-p}{p}}\right) b^{2\frac{p-1}{p}} \leq \frac{1}{p}\left(ab^{\frac{2-p}{p}}\right)^p + \frac{1}{q} b^{2\frac{p-1}{p}q} = \frac{1}{p}a^p b^{2-p} + \left(1 - \frac{1}{p}\right)b^2. \tag{B.5}$$

Applying this inequality with $p = p_x$,

$$\begin{aligned} a &:= (\zeta_x p_x)^{-1/2}\|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\|, \\ b &:= (\zeta_x p_x)^{1/2}\|y^{i+1} - \widehat{y}\|, \end{aligned}$$

for any $\zeta_x > 0$ to the last term of (B.4), we arrive at the estimate

$$\begin{aligned} D_x \geq \theta_x \|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\|^{p_x} - \frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 \\ - \frac{\|y^{i+1} - \widehat{y}\|^{2-p_x}}{p_x^{p_x}\omega_i \zeta_x^{p_x-1}}\|K_y(\widehat{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\widehat{x} - x^{i+1})\|^{p_x} \\ - \frac{p_x - 1}{\omega_i}\zeta_x\|y^{i+1} - \widehat{y}\|^2. \end{aligned}$$

We now use $u^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$ for some $\rho_x, \rho_y \geq 0$, and $\omega_i^{-1} \leq \underline{\omega}^{-1}$ to obtain

$$\theta_x - \|y^{i+1} - \widehat{y}\|^{2-p_x}(p_x^{p_x}\omega_i\zeta_x^{p_x-1})^{-1} \geq \theta_x - \rho_y^{2-p_x}(p_x^{p_x}\underline{\omega}\zeta_x^{p_x-1})^{-1}. \tag{B.6}$$

If $p_x = 1$, we use the assumed inequality $\theta_x \geq \rho_y \underline{\omega}^{-1}$ from (4.8e) to show that the right-hand side of (B.6) is non-negative for any $\zeta_x > 0$. Otherwise we take $\zeta_x := (\underline{\omega}\theta_x p_x^{p_x} \rho_y^{p_x-2})^{1/(1-p_x)}$ to ensure the right-hand side of (B.6) is zero. In either case, $\theta_x - \rho_y^{2-p_x}(p_x^{p_x}\underline{\omega}\zeta_x^{p_x-1})^{-1} \geq 0$ and hence

$$D_x \geq -\frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 - (p_x - 1)\omega_i^{-1}\zeta_x\|y^{i+1} - \widehat{y}\|^2. \tag{B.7}$$

Analogously, from (B.2b) and Cauchy's inequality,

$$\begin{aligned} D_y \geq \theta_y\|K_x(x^i, \widehat{y}) - K_x(x^i, y^i) - K_{xy}(x^i, y^i)(\widehat{y} - y^i)\|^{p_y} - \frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2 \\ - \omega_i\|x^{i+1} - \widehat{x}\|\|K_x(x^i, \widehat{y}) - K_x(x^i, y^i) - K_{xy}(x^i, y^i)(\widehat{y} - y^i)\|. \end{aligned}$$

This has a structure similar to (B.4) with ω_i now as a multiplier. Hence, we apply a similar generalized Young's inequality to the last term with any $\zeta_y > 0$. Noting that $\omega_i \leq \overline{\omega}$, we use the following bound similar to (B.6):

$$\theta_y - \|x^{i+1} - \widehat{x}\|^{2-p_y}\omega_i(p_y^{p_y}\zeta_y^{p_y-1})^{-1} \geq \theta_y - \rho_x^{2-p_y}\overline{\omega}(p_y^{p_y}\zeta_y^{p_y-1})^{-1} \geq 0.$$

The last inequality holds for any $\zeta_y > 0$ if $p_y = 1$ due to the assumed $\theta_y \geq \overline{\omega}\rho_x$ from (4.8d); otherwise, we set $\zeta_y := (\theta_y p_y^{p_y}\rho_x^{p_y-2}\overline{\omega}^{-1})^{1/(1-p_y)}$. We then obtain that

$$D_y \geq -\frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2 - (p_y - 1)\omega_i\zeta_y\|x^{i+1} - \widehat{x}\|^2. \tag{B.8}$$

Combining (4.14), (B.7), and (B.8), we can thus bound

$$\begin{aligned} D &= \eta_i D_x + \eta_{i+1}D_y + \eta_{i+1}D_\omega + \eta_i(\gamma_G - \widetilde{\gamma}_G - \xi_x)\|x^{i+1} - \widehat{x}\|^2 \\ &\quad + \eta_{i+1}(\gamma_{F^*} - \widetilde{\gamma}_{F^*} - \xi_y)\|y^{i+1} - \widehat{y}\|^2 \\ &\geq \eta_{i+1}(\gamma_{F^*} - \widetilde{\gamma}_{F^*} - \xi_y - (p_x - 1)\zeta_x)\|y^{i+1} - \widehat{y}\|^2 - \eta_i\frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 \\ &\quad + \eta_i(\gamma_G - \widetilde{\gamma}_G - \xi_x - (p_y - 1)\zeta_y)\|x^{i+1} - \widehat{x}\|^2 - \eta_{i+1}\frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2 \\ &\quad - \eta_i\frac{L_{yx}}{2}(\omega_i + 2)\rho_y\|x^{i+1} - x^i\|^2 \\ &\geq -\eta_i\frac{\lambda_x + L_{yx}(\omega_i + 2)\rho_y}{2}\|x^{i+1} - x^i\|^2 - \eta_{i+1}\frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2, \end{aligned} \tag{B.9}$$

where in the final step, we have also used (B.3) and the selected ζ_x and ζ_y if $p_x > 1$ or $p_y > 1$ or both. Thus, we obtained exactly the same lower bound as in (4.17). We then continue along the rest of the proof of Theorem 4.2 to obtain the claim. □

It is worth observing that when $p_x \in (1, 2]$ or $p_y \in (1, 2]$, the inequalities (B.3) do not directly bound the respective ρ_y or ρ_x. Hence, we do not need to initalize the corresponding variable locally, unlike when $p_x = 1$ or $p_y = 1$. On the other hand, sufficient strong convexity is required from the corresponding G and F^*.

We start with the lemma ensuring that the iterates stay in the initial neighborhood of the saddle point.

Corollary B.2. *The results of Lemma 5.2 continue to hold if the corresponding conditions of Theorem 4.2 are replaced with those in Corollary B.1.*

Proof. The proof repeats that of Lemma 5.2, applying Corollary B.1 instead of Theorem 4.2 in Step 2. □

We next extend the results of Section 6 to arbitrary choices of both $p_x \in [1,2]$ and $p_y \in [1,2]$. This mainly consists of verifying (B.3A) when $p_y \neq 1$ and (B.3B) when $p_x \neq 1$. Note that it is possible to take $p_x = 1$ and $p_y \neq 1$, or vice versa, as long as the corresponding conditions are satisfied.

Corollary B.3. *The results of Theorem 6.1 continue to hold if Assumption 3.2 (iv) is replaced with Assumption B.1 (iv*) for some $p_x, p_y \in [1,2]$, where in case $p_y \in (1,2]$, (6.1a) is replaced with*

$$\xi_x = \gamma_G - \frac{p_y - 1}{(\theta_y p_y^{p_y} (2\rho_x)^{p_y - 2})^{\frac{1}{p_y - 1}}}, \tag{B.10A}$$

and in case $p_x \in (1,2]$, (6.1b) is replaced with

$$\xi_y = \gamma_{F^*} - \frac{p_x - 1}{(\theta_x p_x^{p_x} (2\rho_y)^{p_x - 2})^{\frac{1}{p_x - 1}}}. \tag{B.10B}$$

Proof. Since conditions (B.10) are sufficient for (B.3) with $\overline{\omega} = \underline{\omega} = 1$ to hold, we can repeat the proof of Theorem 6.1 replacing the references to Theorem 4.2 by references to Corollary B.1 up until (6.5). If $p_x > 1$, we now obtain a lower bound on d_i^x by arguing as in (B.4)–(B.6) with $\widehat{u}$ replaced by $\overline{u}$. Specifically, using (3.5), Assumption B.1 (iv*) at $\overline{u}$, and the generalized Young's inequality (B.5), we obtain for any $\zeta_x > 0$ that

$$\begin{aligned}
d_i^x &\leq -\theta_x \|K_y(\bar{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\bar{x} - x^{i+1})\|^{p_x} \\
&\quad + \|y^{i+1} - \bar{y}\| \|K_y(\bar{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\bar{x} - x^{i+1})\| \\
&\quad + \frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 - \frac{p_y - 1}{(\theta_y p_y^{p_y} (2\rho_x)^{p_y - 2})^{\frac{1}{p_y - 1}}}\|x^{i+1} - \bar{x}\|^2 \\
&\leq \left(\frac{\|y^{i+1} - \bar{y}\|^{2 - p_x}}{p_x^{p_x} \zeta_x^{p_x - 1}} - \theta_x \right) \|K_y(\bar{x}, y^{i+1}) - K_y(x^{i+1}, y^{i+1}) - K_{yx}(x^{i+1}, y^{i+1})(\bar{x} - x^{i+1})\|^{p_x} \\
&\quad + (p_x - 1)\zeta_x \|y^{i+1} - \bar{y}\|^2 + \frac{\lambda_x}{2}\|x^{i+1} - x^i\|^2 - \frac{p_y - 1}{(\theta_y p_y^{p_y} (2\rho_x)^{p_y - 2})^{\frac{1}{p_y - 1}}}\|x^{i+1} - \bar{x}\|^2.
\end{aligned}$$

Inserting $\zeta_x = (\theta_x p_x^{p_x} (2\rho_y)^{p_x - 2})^{1/(1 - p_x)}$ and $\|y^{i+1} - \bar{y}\| \leq 2\rho_y$, we eliminate the first term on the right-hand side. Likewise, if $p_y > 1$, similar steps applied to d_i^y result in

$$d_i^y \leq (p_y - 1)\zeta_y \|x^{i+1} - \bar{x}\|^2 + \frac{\lambda_y}{2}\|y^{i+1} - y^i\|^2 - \frac{p_x - 1}{(\theta_x p_x^{p_x} (2\rho_y)^{p_x - 2})^{\frac{1}{p_x - 1}}}\|y^{i+1} - \bar{y}\|^2$$

for $\zeta_y = (\theta_y p_y^{p_y} (2\rho_x)^{p_y - 2})^{1/(p_y - 1)}$. Using $\|u^{i+1} - u^i\| \to 0$ and the selection of ζ_x and ζ_y, we then obtain the desired estimate $\limsup_{i \to \infty} q_i := \limsup_{i \to \infty} (d_i^x + d_i^y + O(\|u^{i+1} - u^i\|)) \leq 0$. □

Corollary B.4. *The results of Theorem 6.3 continue to hold if Assumption 3.2 (iv) is replaced with Assumption B.1 (iv*) for some $p_x, p_y \in [1,2]$, where in case $p_y \in (1,2]$, (6.6a) is replaced for some $\widetilde{\gamma}_G > 0$ with*

$$\xi_x = \gamma_G - \widetilde{\gamma}_G - \frac{p_y - 1}{(\theta_y p_y^{p_y} (\rho_x)^{p_y - 2})^{\frac{1}{p_y - 1}}}, \tag{B.11A}$$

and in case $p_x \in (1,2]$, (6.6b) is replaced with

$$\xi_y = \gamma_{F^*} - \frac{p_x - 1}{(\theta_x p_x^{p_x} (\rho_y)^{p_x - 2})^{\frac{1}{p_x - 1}}}. \tag{B.11B}$$

Proof. Conditions (B.11) are sufficient for (B.3) with $\overline{\omega} = \underline{\omega} = 1$ to hold; therefore, we can repeat the proof of Theorem 6.3 replacing the references to Theorem 4.2 by references to Corollary B.1. □

Corollary B.5. *The results of Theorem 6.4 continue to hold if Assumption 3.2 (iv) is replaced with Assumption B.1 (iv*) for some $p_x, p_y \in [1,2]$, where in case $p_y \in (1,2]$, (6.9a) is replaced for some $\widetilde{\gamma}_G > 0$ with*

$$\xi_x = \gamma_G - \widetilde{\gamma}_G - \frac{p_y - 1}{(\theta_y p_y^{p_y} (\rho_x)^{p_y - 2} \omega^{-1})^{\frac{1}{p_y - 1}}}, \tag{B.12A}$$

and in case $p_x \in (1,2]$, (6.9b) is replaced for some $\widetilde{\gamma}_{F^*} > 0$ with

$$\xi_y = \gamma_{F^*} - \widetilde{\gamma}_{F^*} - \frac{p_x - 1}{(\omega \theta_x p_x^{p_x} (\rho_y)^{p_x - 2})^{\frac{1}{p_x - 1}}}. \tag{B.12B}$$

Proof. Conditions (B.12) are sufficient for (B.3) with $\overline{\omega} = \underline{\omega} = \omega$ to hold; therefore, we can repeat the proof of Theorem 6.4 replacing the references to Theorem 4.2 by references to Corollary B.1. □

Corollary B.6. *The results of Proposition 6.6 continue to hold if the corresponding conditions of Theorem 6.1, 6.3, or 6.4 are replaced with those in Corollary B.3, B.4, or B.5.*

Proof. The proof repeats that of Proposition 6.6. □

APPENDIX C VERIFICATION OF CONDITIONS FOR STEP FUNCTION PRESENTATION AND POTTS MODEL

Throughout this section, we set $\rho(t) := 2t - t^2$ and $\kappa(x,y) := \rho(\langle x, y \rangle)$ for $x, y \in \mathbb{R}^m$. Then $\rho'(t) = 2(1-t)$ so that

$$\kappa_x(x,y) = 2y(1 - \langle y, x \rangle) \quad \text{and} \quad \kappa_{xy}(x,y) = 2(I - \langle y, x \rangle I - y \otimes x), \tag{C.1A}$$

$$\kappa_y(x,y) = 2x(1 - \langle x, y \rangle) \quad \text{and} \quad \kappa_{yx}(x,y) = 2(I - \langle x, y \rangle I - x \otimes y), \tag{C.1B}$$

where $a \otimes b \in \mathbb{R}^{n \times n}$ is the tensor product between two vectors a and b, producing a matrix of all the combinations of products between the entries.

The following lemma verifies Assumption 3.2 for $K = \kappa$.

Lemma c.1. *Let $R_K > 2$, and suppose $\widehat{x}, \widehat{y} \in \mathbb{R}^m$ for $m \geq 1$ with*

$$0 \leq \langle \widehat{x}, \widehat{y} \rangle I + \widehat{x} \otimes \widehat{y} \leq 2I. \tag{c.2}$$

Then the function $K = \kappa$ defined above satisfies Assumption 3.2 for some $\theta_x, \theta_y > 0$ and some $\rho_x, \rho_y > 0$ dependent on R_K with

$$L_x(y) = 2|y|_2^2, \qquad L_y(x) = 2|x|_2^2, \qquad L_{yx} = 4(|\widehat{y}|_2 + \rho_y),$$

as well as the constants $\xi_x, \xi_y \in \mathbb{R}$, $\lambda_x, \lambda_y \geq 0$ satisfying $\lambda_x \xi_x > 2(\lambda_x + |\widehat{y}|_2^2)|\widehat{y}|_2^2$, $\xi_y > 0$, and $\lambda_y > |\widehat{x}|_2^2$.

Proof. First, Assumption 3.2 (i) holds everywhere since $K \in C^\infty(\mathbb{R}^m)$. To verify Assumption 3.2 (ii), we observe using (c.1) that

$$\kappa_x(x', y) - \kappa_x(x, y) = 2(y \otimes y)(x - x'), \tag{c.3a}$$

$$\kappa_{xy}(x, y') - \kappa_{xy}(x, y) = 2\langle y - y', x \rangle I + 2(y - y') \otimes x, \tag{c.3b}$$

$$\kappa_y(x, y') - \kappa_y(x, y) = 2(x \otimes x)(y - y'), \tag{c.3c}$$

$$\kappa_{yx}(x', y) - \kappa_{yx}(x, y) = 2\langle x - x', y \rangle I + 2(x - x') \otimes y. \tag{c.3d}$$

Hence L_x, L_y, and L_{yx} are as claimed.

To verify Assumption 3.2 (iii), we first of all observe using (c.2) that

$$|\kappa_{xy}(\widehat{x}, \widehat{y})|_2 = 2|I - \langle \widehat{y}, \widehat{x} \rangle I - \widehat{y} \otimes \widehat{x}|_2 \leq 2.$$

Therefore $\sup_{(x,y) \in \mathbb{B}(\widehat{x}, \rho_x) \times \mathbb{B}(\widehat{y}, \rho_y)} |\kappa_{xy}(x, y)|_2 \leq R_K$ for some $\rho_x, \rho_y > 0$ dependent on $R_K > 2$.

Finally, to verify Assumption 3.2 (iv), we start with (3.4a), i.e.,

$$\begin{aligned}\langle \kappa_x(x', \widehat{y}) - \kappa_x(\widehat{x}, \widehat{y}), x - \widehat{x} \rangle + \xi_x |x - \widehat{x}|_2^2 \\ \geq \theta_x |\kappa_y(\widehat{x}, y) - \kappa_y(x, y) - \kappa_{yx}(x, y)(\widehat{x} - x)|_2 - \frac{\lambda_x}{2} |x - x'|_2^2.\end{aligned}$$

Expanding the equation using (c.1), (c.3), and

$$\begin{aligned}&\kappa_y(\widehat{x}, y) - \kappa_y(x, y) - \kappa_{yx}(x, y)(\widehat{x} - x) \\ &\quad = 2\widehat{x}(1 - \langle \widehat{x}, y \rangle) - 2x(1 - \langle x, y \rangle) - 2(I - \langle x, y \rangle I - x \otimes y)(\widehat{x} - x) \\ &\quad = 2[\langle x, y \rangle x - \langle \widehat{x}, y \rangle \widehat{x} + (\langle x, y \rangle I + x \otimes y)(\widehat{x} - x)] \\ &\quad = 2[\langle x - \widehat{x}, y \rangle \widehat{x} + (x \otimes y)(\widehat{x} - x)] \\ &\quad = -2((\widehat{x} - x) \otimes y)(\widehat{x} - x),\end{aligned}$$

we require that

$$2\langle \widehat{x} - x', x - \widehat{x} \rangle_{\widehat{y} \otimes \widehat{y}} + \xi_x |x - \widehat{x}|_2^2 \geq 2\theta_x |y|_2 |x - \widehat{x}|_2^2 - \frac{\lambda_x}{2} |x - x'|_2^2. \tag{c.4}$$

Taking any $\alpha > 0$, this will hold by Cauchy's and Young's inequalities if $\xi_x \geq (2+\alpha)|\widehat{y}|_2^2 + 2\theta_x |y|_2$ and $\lambda_x / 2 \geq \alpha^{-1} |\widehat{y}|_2^2$. If $|\widehat{y}|_2 = 0$, clearly these hold for some $\alpha, \theta_x > 0$. Otherwise, solving α from

the latter as an equality, i.e., taking $\alpha = 2\lambda_x^{-1}|\widehat{y}|_2^2$, the former holds if $\xi_x \geq 2(1 + \lambda_x^{-1}|\widehat{y}|_2^2)|\widehat{y}|_2^2 + 2\theta_x|y|_2$. If $\lambda_x\xi_x > 2(\lambda_x + |\widehat{y}|_2^2)|\widehat{y}|_2^2$, this holds for some $\theta_x, \rho_x, \rho_y > 0$ in a neighborhood $\mathbb{B}(\widehat{x}, \rho_x) \times \mathbb{B}(\widehat{y}, \rho_y)$ of $(\widehat{x}, \widehat{y})$.

It remains to verify (3.4b), i.e.,

$$\langle \kappa_y(x,y) - \kappa_y(x,y') + \kappa_y(\widehat{x},\widehat{y}) - \kappa_y(\widehat{x},y), y - \widehat{y}\rangle + \xi_y|y-\widehat{y}|_2^2 \\ \geq \theta_y|\kappa_x(x',\widehat{y}) - \kappa_x(x',y') - \kappa_{xy}(x',y')(\widehat{y}-y')|_2 - \frac{\lambda_y}{2}|y-y'|_2^2.$$

Again, using (C.1) and (C.3) we expand this as

$$2\langle y'-y, y-\widehat{y}\rangle_{x\otimes x} + 2|y-\widehat{y}|_{\widehat{x}\otimes\widehat{x}}^2 + \xi_y|y-\widehat{y}|_2^2 \geq 2\theta_y|x'|_2|y'-\widehat{y}|_2^2 - \frac{\lambda_y}{2}|y-y'|_2^2.$$

Rearranging the θ_y-term, we see that this holds if

$$2\langle y'-y, y-\widehat{y}\rangle_{x\otimes x - 2\theta_y|x'|_2 I} + 2|y-\widehat{y}|_{\widehat{x}\otimes\widehat{x}}^2 + (\xi_y - 2\theta_y)|x'|_2|y-\widehat{y}|_2^2 \\ \geq \left(2\theta_y|x'|_2 - \frac{\lambda_y}{2}\right)|y'-y|_2^2.$$

Rearranging and estimating the first term as

$$2\langle y'-y, y-\widehat{y}\rangle_{x\otimes x - 2\theta_y|x'|_2 I} = 2\langle y'-y, x\rangle\langle y-\widehat{y}, x\rangle - 4\theta_y|x'|_2\langle y'-y, y-\widehat{y}\rangle \\ \geq -2|y-\widehat{y}|_{x\otimes x}^2 - \frac{1}{2}|y'-y|_{x\otimes x}^2 - 4\theta_y|x'|_2|y'-y|_2^2 - \theta_y|x'|_2|y-\widehat{y}|_2^2$$

and then using Young's inequality on both parts, we obtain the condition

$$2\left(|y-\widehat{y}|_{\widehat{x}\otimes\widehat{x}}^2 - |y-\widehat{y}|_{x\otimes x}^2\right) + (\xi_y - 3\theta_y)|x'|_2|y-\widehat{y}|_2^2 \geq \left(\frac{1}{2}|x|_2^2 + 6\theta_y|x'|_2 - \frac{\lambda_y}{2}\right)|y'-y|_2^2.$$

If $\xi_y > 0$ and $\lambda_y > |\widehat{x}|_2^2$, this holds for some $\theta_y, \rho_y, \rho_x > 0$ in $\mathbb{B}(\widehat{x}, \rho_x) \times \mathbb{B}(\widehat{y}, \rho_y)$. □

We comment on the condition (C.2) on the primal–dual solutions pair $\widehat{x}, \widehat{y} \in \mathbb{R}$. First, for $m = 1$, this condition reduces to $\widehat{x}\widehat{y} \in [0,1]$. This is necessarily satisfied in the case of the step function (where $f^* = \delta_{[0,\infty)}$) and in the case of the ℓ^0 function (where $f^* = 0$) as in both cases, $\widehat{x}\widehat{y} \in \{0,1\}$ by the dual optimality condition $\kappa_y(\widehat{x},\widehat{y}) \in \partial f^*(\widehat{y})$. Furthermore, if we take $f_\gamma^* = \frac{\gamma}{2}|\cdot|_2^2$ for some $\gamma \geq 0$, then for any $m \geq 1$ the dual optimality condition reads $2\widehat{x}(1 - \langle\widehat{x},\widehat{y}\rangle) = \gamma\widehat{y}$, i.e, $\widehat{y} = 2\widehat{x}(\gamma + 2|\widehat{x}|_2^2)^{-1}$, for which (C.2) is easily verified.

The following lemma shows that Assumption 3.2 remains valid if we include a linear operator in the primal component.

Lemma C.2. *Let $K(x,y) = \widetilde{K}(Ax,y)$ for some $A \in \mathbb{L}(X;Z)$ and $\widetilde{K} \in C^1(Z\times Y)$ on Hilbert spaces X, Y, Z. Suppose $\widetilde{K}$ satisfies Assumption 3.2 at $(\widehat{z},\widehat{y}) := (A\widehat{x},\widehat{y})$. Mark the corresponding constants with a tilde: $\widetilde{L}_z$, $\widetilde{R}_K$, and so on. Then K satisfies Assumption 3.2 with $R_K := \widetilde{R}_K\|A\|$; $\xi_x = \|A\|\widetilde{\xi}_z$, $\xi_y = \widetilde{\xi}_y$; $\lambda_x = \|A\|\widetilde{\lambda}_z$, $\lambda_y = \widetilde{\lambda}_y$; $\theta_x = \widetilde{\theta}_z$, $\theta_y = \widetilde{\theta}_y\|A\|^{-1}$; $\rho_x = \|A\|^{-1}\widetilde{\rho}_x$, and $\rho_y = \widetilde{\rho}_y$ as well as*

$$L_x(y) = \|A\|^2\widetilde{L}_z(y), \qquad L_y(x) = \widetilde{L}_y(Ax), \qquad L_{yx} = \|A\|^2\widetilde{L}_{yz}. \tag{C.5}$$

Proof. Observe first of all that by the chain rule,

$$K_x(x,y) = A^*\widetilde{K}_z(Ax,y), \qquad K_y(x,y) = \widetilde{K}_y(Ax,y), \qquad K_{xy}(x,y) = A^*\widetilde{K}_{zy}(Ax,y),$$

and hence Assumption 3.2 (i) holds for K if it holds for $\widetilde{K}$.

Let now Assumption 3.2 (ii) hold for $\widetilde{K}$ with $\widetilde{L}_x$, $\widetilde{L}_y$, and $\widetilde{L}_{yx}$. Observing that

$$A\mathbb{B}(\widehat{x},\rho_x) \times \mathbb{B}(\widehat{y},\rho_y) \subset \mathbb{B}(\widehat{z},\widetilde{\rho}_x) \times \mathbb{B}(\widehat{y},\widetilde{\rho}_y), \tag{C.6}$$

Assumption 3.2 (ii) thus also holds with the function of (C.5). Similarly in Assumption 3.2 (iii), we can take $R_K := \widetilde{R}_K\|A\|$.

Finally, we expand Assumption 3.2 (iv) for K as

$$\begin{aligned}
&\langle \widetilde{K}_z(z',\widehat{y}) - \widetilde{K}_z(\widehat{z},\widehat{y}), z-\widehat{z}\rangle + \xi_x\|x-\widehat{x}\|^2 \\
&\quad \geq \theta_x\|\widetilde{K}_y(\widehat{z},y) - \widetilde{K}_y(z,y) - \widetilde{K}_{yz}(z,y)(\widehat{z}-z)\| - \frac{\lambda_x}{2}\|x-x'\|^2
\end{aligned}$$

and

$$\begin{aligned}
&\langle \widetilde{K}_y(z,y) - \widetilde{K}_y(z,y') + \widetilde{K}_y(\widehat{z},\widehat{y}) - \widetilde{K}_y(\widehat{z},y), y-\widehat{y}\rangle + \xi_y\|y-\widehat{y}\|^2 \\
&\quad \geq \theta_y\|A^*[\widetilde{K}_z(z',\widehat{y}) - \widetilde{K}_z(z',y') - \widetilde{K}_{zy}(z',y')(\widehat{y}-y')]\| - \frac{\lambda_y}{2}\|y-y'\|^2,
\end{aligned}$$

where $z = Ax$, $z' = Ax'$, and $\widehat{z} = A\widehat{x}$. Since $\|z-z'\| \leq \|A\|\|x-x'\|$, etc., this follows from Assumption 3.2 (iv) for $\widetilde{K}$ with the constants as claimed. □

Applying this lemma to $\widetilde{K}(z,y) = \sum_{k=1}^n \kappa(z_k,y_k)$, we can thus lift the scalar estimates for $K = \kappa$ as in (C.1) to the corresponding estimates on $K(x,y) := \sum_{k=1}^n \kappa([D_h x]_k, y_k)$ as used in the Potts model example.

REFERENCES

[1] Aragón Artacho & Geoffroy, Metric subregularity of the convex subdifferential in Banach spaces, *J. Nonlinear Convex Anal.* 15 (2014), 35–47, arXiv: 1303.3654.

[2] Aragón Artacho & Geoffroy, Characterization of metric regularity of subdifferentials, *Journal of Convex Analysis* 15 (2008), 365–380.

[3] Attouch, Bolte & Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods, *Mathematical Programming* 137 (2013), 91–129, DOI: 10.1007/s10107-011-0484-9.

[4] Bauschke & Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2017, DOI: 10.1007/978-3-319-48311-5.

[5] Benning, Knoll, Schönlieb & Valkonen, Preconditioned ADMM with nonlinear operator constraint, in: *System Modeling and Optimization: 27th IFIP TC 7 Conference, CSMO 2015, Sophia Antipolis, France, June 29–July 3, 2015, Revised Selected Papers*, ed. by Bociu, Désidéri & Habbal, Springer International Publishing, 2016, 117–126, DOI: 10.1007/978-3-319-55795-3_10, arXiv: 1511.00425,

[6] Borzì & Kanzow, Formulation and numerical solution of Nash equilibrium multiobjective elliptic control problems, *SIAM Journal on Control and Optimization* 51 (2013), 718–744, DOI: 10.1137/120864921.

[7] Chambolle, An algorithm for total variation minimization and applications, *Journal of Mathematical Imaging and Vision* 20 (2004), 89–97, DOI: 10.1023/b:jmiv.0000011325.36760.1e.

[8] Chambolle & Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision* 40 (2011), 120–145, DOI: 10.1007/s10851-010-0251-1.

[9] Clason & Kunisch, A convex analysis approach to multi-material topology optimization, *ESAIM: Mathematical Modelling and Numerical Analysis* 50 (2016), 1917–1936, DOI: 10.1051/m2an/2016012.

[10] Clason, Mazurenko & Valkonen, Acceleration and global convergence of a first-order primal-dual method for nonconvex problems, *SIAM Journal on Optimization* 29 (2019), 933–963, DOI: 10.1137/18m1170194, ARXIV: 1802.03347,

[11] Clason, Mazurenko & Valkonen, Julia codes for "Primal–dual proximal splitting and generalized conjugation in non-smooth non-convex optimization", Online resource on Zenodo, 2020, DOI: 10.5281/zenodo.3647614.

[12] Clason & Valkonen, Primal-dual extragradient methods for nonlinear nonsmooth PDE-constrained optimization, *SIAM Journal on Optimization* 27 (2017), 1313–1339, DOI: 10.1137/16m1080859, ARXIV: 1606.06219,

[13] Drori, Sabach & Teboulle, A simple algorithm for a class of nonsmooth convex–concave saddle-point problems, *Operations Research Letters* 43 (2015), 209–214, DOI: 10.1016/j.orl.2015.02.001.

[14] Ekeland & Temam, *Convex analysis and variational problems*, SIAM, 1999.

[15] Elster & Wolf, Recent results on generalized conjugate functions, in: *Trends in Mathematical Optimization: 4th French-German Conference on Optimization*, ed. by Hoffmann, Zowe, Hiriart-Urruty & Lemarechal, Birkhäuser Basel, 1988, 67–78, DOI: 10.1007/978-3-0348-9297-1_5.

[16] Facchinei & Kanzow, Generalized Nash equilibrium problems, *Ann. Oper. Res.* 175 (2010), 177–211, DOI: 10.1007/s10479-009-0653-x.

[17] Flåm & Antipin, Equilibrium programming using proximal-like algorithms, *Math. Programming* 78 (1997), 29–41, DOI: 10.1007/bf02614504.

[18] Geman & Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1984), 721–741, DOI: 10.1109/tpami.1984.4767596.

[19] Hamedani & Aybat, A primal-dual algorithm for general convex-concave saddle point problems, 2018, ARXIV: 1803.01401.

[20] He, Juditsky & Nemirovski, Mirror Prox algorithm for multi-term composite minimization and semi-separable problems, *Computational Optimization and Applications* 61 (2015), 275–319, DOI: 10.1007/s10589-014-9723-3.

[21] He & Monteiro, An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems, *SIAM Journal on Optimization* 26 (2016), 29–56, DOI: 10.1137/14096757x.

[22] Juditsky & Nemirovski, First order methods for nonsmooth convex large-scale optimization, I: general purpose methods, in: *Optimization for Machine Learning*, ed. by Sra, Nowozin & Wright, MIT Press, 2011, 121–148, URL: www.jstor.org/stable/j.ctt5hhgpg.9.

[23] Juditsky & Nemirovski, First order methods for nonsmooth convex large-scale optimization, II: utilizing problems structure, in: *Optimization for Machine Learning*, ed. by Sra, Nowozin & Wright, MIT Press, 2011, 149–183, URL: www.jstor.org/stable/j.ctt5hhgpg.10.

[24] Kolossoski & Monteiro, An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems, *Optimization Methods and Software* 32 (2017), 1244–1272, DOI: 10.1080/10556788.2016.1266355.

[25] Krawczyk & Uryasev, Relaxation algorithms to find Nash equilibria with economic applications, *Environmental Modeling & Assessment* 5 (2000), 63–73, DOI: 10.1023/a:1019097208499.

[26] Martinez-Legaz, Generalized convex duality and its economic applications, in: *Handbook of Generalized Convexity and Generalized Monotonicity*, ed. by Hadjisavvas, Komlósi & Schaible, Springer, 2005, 237–292, DOI: 10.1007/0-387-23393-8_6.

[27] Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, *SIAM Journal on Optimization* 15 (2004), 229–251, DOI: 10.1137/s1052623403425629.

[28] Nesterov, Smooth minimization of non-smooth functions, *Mathematical Programming* 103 (2005), 127–152, DOI: 10.1007/s10107-004-0552-5.

[29] Nikaidô & Isoda, Note on non-cooperative convex games, *Pacific J. Math.* 5 (1955), 807–815, DOI: 10.2140/pjm.1955.5.807,

[30] Rasband, ImageJ, U. S. National Institutes of Health, Bethesday, Maryland, USA, 1997–2018, URL: imagej.nih.gov/ij/.

[31] Rosen, Existence and uniqueness of equilibrium points for concave n-person games, *Econometrica* 33 (1965), 520–534, DOI: 10.2307/1911749.

[32] Singer, *Duality for Nonconvex Approximation and Optimization*, Springer-Verlag New York, 2006, DOI: 10.1007/0-387-28395-1.

[33] Storath, Weinmann & Demaret, Jump-sparse and sparse recovery using Potts functionals, *IEEE Transactions on Signal Processing* 62 (2014), 3654–3666, DOI: 10.1109/tsp.2014.2329263.

[34] Storath, Weinmann, Frikel & Unser, Joint image reconstruction and segmentation using the Potts model, *Inverse Problems* 31 (2015), 025003, DOI: 10.1088/0266-5611/31/2/025003.

[35] Valkonen, A primal-dual hybrid gradient method for non-linear operators with applications to MRI, *Inverse Problems* 30 (2014), 055012, DOI: 10.1088/0266-5611/30/5/055012, ARXIV: 1309.5032,

[36] Valkonen, Testing and non-linear preconditioning of the proximal point method, *Applied Mathematics and Optimization* (2018), DOI: 10.1007/s00245-018-9541-6, ARXIV: 1703.05705,

[37] Valkonen & Pock, Acceleration of the PDHGM on partially strongly convex functions, *Journal of Mathematical Imaging and Vision* 59 (2017), 394–414, DOI: 10.1007/s10851-016-0692-2, ARXIV: 1511.06566,

[38] Von Heusinger & Kanzow, Optimization reformulations of the generalized Nash equilibrium problem using Nikaido-Isoda-type functions, *Comput. Optim. Appl.* 43 (2009), 353–377, DOI: 10.1007/s10589-007-9145-6.