First-order primal-dual methods for nonsmooth non-convex optimisation

Tuomo Valkonen

Abstract We provide an overview of primal-dual algorithms for nonsmooth and non-convex-concave saddle-point problems. This flows around a new analysis of such methods, using Bregman divergences to formulate simplified conditions for convergence.

1 Introduction

Interesting imaging problems can often be written in the general form

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

where X and Y are Banach spaces, $K\in C^1(X,Y)$, and $F: X\to\overline{\mathbb{R}}$ and $G_*: Y\to\overline{\mathbb{R}}$ are convex, proper, lower semicontinuous functions with G_* the preconjugate of some $G: Y^*\to\overline{\mathbb{R}}$, meaning $G=(G_*)^*$. The functions F and G_* may be nonsmooth.

A common instance of (S) is when $K(x,y)=\langle Ax|y\rangle$ for a linear operator $A\in\mathbb{L}(X;Y^*)$ with $\langle\,\cdot\,|\,\cdot\,\rangle: Y^*\times Y\to\mathbb{R}$ denoting the dual product. Then (S) arises from taking (pre)conjugates in

$$\min_{x\in X}\ F(x)+G(Ax). \tag{1}$$

Optimisation problems of this type can effectively model linear *inverse problems*; typically one would attempt to minimise the sum of a data-term and a regulariser,

$$\min_{x\in X}\ \Phi(z-Tx)+G(Ax), \tag{2}$$

where

Tuomo Valkonen
Center for Mathematical Modeling, Escuela Politécnica Nacional, Quito, Ecuador *and* Department of Mathematics and Statistics, University of Helsinki, Finland; e-mail: tuomo.valkonen@iki.fi

- $T :\in \mathbb{L}(X; \mathbb{R}^n)$ is a forward operator, mapping our unknown x into a finite number of measurements. Simple examples include blurring, subsampling, the Fourier and Radon transforms, and combinations thereof.
- Φ models noise ν in the data $z \in \mathbb{R}^n$; for normal-distributed noise, $\Phi(z) = \frac{1}{2}\|z\|^2$;
- $G \circ A$ is a typically nonsmooth regularisation term that models our prior assumptions on what a good solution to the ill-posed problem $z = Tx + \nu$ should be; in imaging, what "looks good". For conventional total variation regularisation on a domain $\Omega \subset \mathbb{R}^m$ one would take $G(y^*) = \alpha\|y^*\|_{\mathcal{M}(\Omega;\mathbb{R}^m)}$ the Radon norm of the measure y^* weighted by the regularisation parameter $\alpha > 0$, and $A = D \in \mathbb{L}(\mathrm{BV}(\Omega); \mathcal{M}(\Omega; \mathbb{R}^m))$ the distributional derivative.

To work with a nonlinear forward operator T, if $\Phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous, we can write (2) using the conjugate Φ^* as

$$\min_{x\in X} \max_{(y_1,y_2)\in\mathbb{R}^n\times Y} K(x,(y_1,y_2)) - \Phi^*(y_1) - G_*(y_2) \tag{3}$$

for $K(x,(y_1,y_2)) = \langle z - T(x)|y_1\rangle + \langle Ax|y_2\rangle$. Examples include diffusion tensor imaging, phase and amplitude reconstruction [37], as well as inverse problems governed by nonlinear partial differential equations [19, 17].

If we take $K(x,y) = E(x) + \langle Ax|y\rangle$ for some $E \in C^1(X)$, (S) arises from

$$\min_{x\in X} F(x) + E(x) + G(Ax).$$

The distinction between F and E is important in the algorithms that we study, as they perform a gradient step with respect to E and a proximal step with respect to F.

Finally, fully general K in (S) was shown in [18] to be useful for splitting highly nonsmooth and nonconvex G in (1) into a convex and possibly nonsmooth *generalised preconjugate* G_* and a smooth kernel K. For example, $t \mapsto |t|_0 :=$ (0 if $t = 0$ and 1 otherwise) $= \sup_{s\in\mathbb{R}} \rho(st)$ for $\rho(t) = 2t - t^2$. This is useful with the Potts segmentation model. Also the step function can be written $\chi_{(0,\infty)}(t) = \sup_{s\geq 0} \rho(st)$.

We introduce in Section 3 methods for (S) inspired by the *primal-dual proximal splitting* (PDPS) of [12, 35] for bilinear K, commonly known as the *Chambolle–Pock method*. We work in Banach spaces, as was done in [29]. To do this, in Section 2, we introduce and recall the crucial properties of so-called *Bregman divergences*.

Our main reason for working with Bregman divergences is, however, not the generality of Banach spaces. Rather, they provide a powerful proof tool to deal with the general K in (S). This approach allows us in Section 4 to significantly simplify and better explain the original proofs and conditions of [12, 37, 17, 18, 32]. Without additional effort, they also allow us to present block-adapted methods like those in [41, 40, 32]. The three main ingredients that ensure convergence are:

(i) A three-point identity, satisfied by Bregman divergences (Section 4.1),
(ii) Ellipticity of the algorithm-defining Bregman divergences (Sections 4.2 and 4.3), and
(iii) A non-smooth second-order growth condition (Sections 4.4 and 4.5).

In the present overview, with focus on key concepts and aiming to avoid technical complications, we only cover, weak, strong, and linear convergence of iterates, and the convergence of gap functionals in the convex–concave case.

In Section 5 we improve the basic method by adding dependencies to earlier iterates, a form inertia. This is needed to develop an an effective algorithm for K not affine in y. We finish in Section 6 with pointers to alternative methods and further extensions.

2 Bregman divergences

The norm and inner product in a Hilbert space X satisfy the three-point identity

$$\langle x-y, x-z\rangle_X = \frac{1}{2}\|x-y\|_X^2 - \frac{1}{2}\|y-z\|_X^2 + \frac{1}{2}\|x-z\|_X^2 \quad (x,y,z\in X). \tag{4}$$

This is crucial for convergence proofs of optimisation methods [39], so we would like to have something similar in Banach spaces—or other more general spaces. Towards this end, we let $J : X \to \mathbb{R}$ be a Gâteaux-differentiable function.[1] Then one can define the asymmetric *Bregman divergence*

$$B_J(z,x) := J(z) - J(x) - \langle DJ(x)|z-x\rangle_X \quad (x,z\in X). \tag{5}$$

This function is non-negative *if and only if*[2] the *generating function* J is convex; it is not in general a true distance, as it can happen that $B_J(x,z) = 0$ although $x = z$.

Writing D_1 for the Gâteaux derivative with respect to the first parameter, the Bregman divergence satisfies for any $\bar x \in X$ the *three-point identity*

$$\begin{aligned}\langle D_1 B_J(x,z)|x-\bar x\rangle_X &= \langle DJ(x) - DJ(z)|x-\bar x\rangle_X \\ &= B_J(\bar x, x) - B_J(\bar x, z) + B_J(x,z).\end{aligned} \tag{6}$$

Indeed, writing the right-hand side out, we have

$$\begin{aligned}B_J(\bar x,x) - B_J(\bar x,z) + B_J(x,z) = {} & [J(\bar x) - J(x) - \langle DJ(x)|\bar x - x\rangle_X] \\ & - [J(\hat x) - J(z) - \langle DJ(z)|\hat x - z\rangle_X] \\ & + [J(x) - J(z) - \langle DJ(z)|x-z\rangle_X],\end{aligned}$$

which immediately gives the three-point identity.

Example 1 In a Hilbert space X, the *standard generating function* $J = N_X := \frac{1}{2}\|\cdot\|_X^2$ yields $B_J(z,x) = \frac{1}{2}\|z-x\|_X^2$, so (6) recovers (4).

[1] The differentiability assumption is for notational and presentational simplicity; otherwise we would need to write the Bregman divergence as $B_J^p(z,x) := J(z) - J(x) - \langle p|z-x\rangle_X$ for some subdifferential p of J, and define explicit updates of this subdifferential in algorithms.

[2] For the entirely algebraic proof of the “only if”, see [28, Theorem 4.1.1].

We will frequently require B_J to be *non-negative* or *semi-elliptic* ($\gamma = 0$) or *elliptic* ($\gamma > 0$) within some $\Omega \subset X$. These notions mean that

$$B_J(z, x) \geq \frac{\gamma}{2}\|z - x\|_X^2 \quad (x, z \in \Omega). \tag{7}$$

Equivalently, this defines J to be *(γ-strongly) subdifferentiable* within Ω. When $\Omega = X$, we simply call B_J (semi-)elliptic and J (γ-strongly) subdifferentiable.[3]

We will in Section 5 also need a Cauchy inequality for Bregman divergences. We base this on strong subdifferentiability and the smoothness property (8) in the next lemma. The latter holding with $\Omega = X$ implies that DJ is L-Lipschitz, and in Hilbert spaces is equivalent to this property; see [3, Theorem 18.15] or [39, Appendix C].

Lemma 1 *Suppose $J : X \to \mathbb{R}$ is Gâteaux-differentiable and γ-strongly subdifferentiable within Ω, and satisfies for some $L > 0$ the* subdifferential smoothness

$$\frac{1}{2L}\|DJ(x) - DJ(y)\|_{X^*}^2 \leq J(x) - J(y) - \langle DJ(y)|x - y\rangle \quad (x, y \in \Omega). \tag{8}$$

Then, for any $\alpha > 0$,

$$|\langle D_1 B_J(x, y)|z - x\rangle| \leq \frac{L}{\alpha}B_J(x, y) + \frac{\alpha}{\gamma}B_J(z, x) \quad (x, y, z \in \Omega).$$

Proof By Cauchy's inequality,

$$|\langle D_1 B_J(x, y)|z - x\rangle| \leq \frac{1}{2\alpha}\|DJ(x) - DJ(y)\|_{X^*}^2 + \frac{\alpha}{2}\|z - x\|_X^2.$$

By the strong convexity, $\frac{\gamma}{2}\|z - x\|_X^2 \leq B_J(z, x)$, and by the smoothness property (8), $\frac{1}{2L}\|DJ(x) - DJ(y)\|_{X^*}^2 \leq B_J(x, y)$. Together these estimates yield the claim. □

3 Primal-dual proximal splitting

We now formulate a basic version of our primal-dual method. Later in Section 5 we improve it be more effective when K is not affine in y.

Notation

Throughout the manuscript, we combine the primal and dual variables x and y into variables involving the letter u:

$$u = (x, y), \quad u^k = (x^k, y^k), \quad \hat{u} = (\hat{x}, \hat{y}), \quad \text{etc.}$$

[3] In Banach spaces strong subdifferentiability is implied by strong convexity, as defined without subdifferentials. In Hilbert spaces the two properties are equivalent.

3.1 Optimality conditions and proximal points

Let the *Lagrangian*

$$\mathcal{L}(x, y) := F(x) + K(x, y) - G_*(y).$$

A *saddle point* $\hat{u} = (\hat{x}, \hat{y})$ of the problem (S) satisfies, by definition

$$\mathcal{L}(\hat{x}, y) \le \mathcal{L}(\hat{x}, \hat{y}) \le \mathcal{L}(x, \hat{y}) \quad \text{for all } u = (x, y) \in X \times Y.$$

Writing $D_x K$ and $D_y K$ for the Gâteaux derivatives of K with respect to the two variables, if K is convex-concave, basic results in convex analysis [24, 3] show that

$$-D_x K(\hat{x}, \hat{y}) \in \partial F(\hat{x}) \quad \text{and} \quad D_y K(\hat{x}, \hat{y}) \in \partial G_*(\hat{y}) \tag{9}$$

is necessary and sufficient for $\hat{u}$ to be saddle point. If K is C^1, the theory of generalised subdifferentials of Clarke [16] still indicates[4] the necessity of (9).

We can alternatively write (9) as

$$0 \in H(\hat{u}) := \begin{pmatrix} \partial F(\hat{x}) + D_x K(\hat{x}, \hat{y}) \\ \partial G_*(\hat{y}) - D_y K(\hat{x}, \hat{y}) \end{pmatrix}. \tag{10}$$

If X and Y were Hilbert spaces, we could in principle use the classical *proximal point method* [33, 36] to solve (10): given step length parameters $\tau_k > 0$, iteratively solve u^{k+1} from

$$0 \in H(u^{k+1}) + \tau_k^{-1}(u^{k+1} - u^k). \tag{11}$$

In practise the steps of the method are too expensive to realise as the primal and dual iterates x^{k+1} and y^{k+1} are coupled: generally, one cannot solve one before the other.

Fortunately, the iterates can be decoupled by introducing a *preconditioner* that switches $D_x K(x^{k+1}, y^{k+1})$ on the first line of $H(u^{k+1})$ to $D_x K(x^k, y^k)$. This gives rise to the *primal-dual proximal splitting* (PDPS), introduced in [12, 35] for bilinear $K(x, y) = \langle Ax|y\rangle$. That the method is actually a preconditioned proximal point method was first observed in [27]. In the following, we describe its extension to general K from [37, 17, 18] and to Banach spaces.

3.2 Algorithm formulation

Given Gâteaux-differentiable functions $J_X : X \to \overline{\mathbb{R}}$ and $J_Y : Y \to \overline{\mathbb{R}}$ with the corresponding Bregman divergences $B_X := B_{J_X}$ and $B_Y := B_{J_Y}$, we define

$$J^0(x, y) := J_X(x) + J_Y(y) - K(x, y). \tag{12}$$

[4] The Fermat-rule $0 \in \partial_C[F + K(\,\cdot\,, \hat{y})](\hat{x})$ holds. Since F is convex and $K(\,\cdot\,, \hat{y})$ is C^1, $\hat{x}$ is a regular point of both, so also the subdifferential sum rule holds. We argue $G_* + K(\hat{y}, \,\cdot\,)$ similarly.

Introducing the short-hand notation $B^0 := B_{J^0}$, we propose to solve (10) through the iterative solution of

$$0 \in H(u^{k+1}) + D_1 B^0(u^{k+1}, u^k) \tag{13}$$

for u^{k+1}. We expand and rearrange this as:

Primal-dual Bregman-proximal splitting (PDBS)

Iteratively over $k \in \mathbb{N}$, solve for x^{k+1} and y^{k+1}:

$$\begin{aligned} DJ_X(x^k) - D_x K(x^k, y^k) &\in DJ_X(x^{k+1}) + \partial F(x^{k+1}) \quad \text{and} \\ DJ_Y(y^k) - D_y K(x^k, y^k) &\in DJ_Y(y^{k+1}) + \partial G_*(y^{k+1}) - 2D_y K(x^{k+1}, y^{k+1}). \end{aligned} \tag{14}$$

If $DJ_X + \tau\partial F$ can be inverted easily, which is to say that F is *prox-simple* with respect to J_X, we readily obtain x^{k+1}. For y^{k+1}, the same is true if K is affine in y and G_* is prox-simple with respect to J_Y. If, however, K is not affine in y, it is practically unlikely that $\partial G_* - 2D_y K(x^{k+1}, \cdot)$ would be prox-simple. We will therefore improve the method for general K in Section 5, after first studying fundamental ideas behind convergence proofs in the following Section 4.

If X and Y are Hilbert spaces with $J_X = \tau^{-1} N_X$ and $J_Y = \sigma^{-1} N_Y$ the standard generating functions divided by some step length parameters $\tau, \sigma > 0$, (14) becomes

Primal–dual proximal splitting (PDPS)

Iterate over $k \in \mathbb{N}$:

$$\begin{aligned} x^{k+1} &:= \mathrm{prox}_{\tau F}(x^k - \tau \nabla_x K(x^k, y^k)), \\ y^{k+1} &:= \mathrm{prox}_{\sigma[G_* - 2K(x^{k+1}, \cdot)]}(y^k - \sigma \nabla_y K(x^k, y^k)). \end{aligned} \tag{15}$$

The proximal map

$$\mathrm{prox}_{\tau F}(x) := (I + \tau\partial F)^{-1}(x) = \underset{\tilde{x} \in X}{\arg\min} \left(\tau F(\tilde{x}) + \frac{1}{2}\|\tilde{x} - x\|_X^2 \right).$$

In finite dimensions, several worked out proximal maps may be found online [15] or in the book [4]. Some extend directly to Hilbert spaces or by superposition to L^2.

Remark 1 For K affine in y, (15) corresponds to the "linearised" variant of the NL-PDPS of [37]. The "exact" variant, studied in further detail in [17], updates

$$y^{k+1} := \mathrm{prox}_{\sigma G_*}(y^k + \sigma \nabla_y K(2x^{k+1} - x^k, y^k)).$$

If K is bilinear the two variants are the exactly same PDPS of [12]. For K not affine in y, the method is neither the generalised PDPS of [18] nor the version for convex-concave K from [26].

3.3 Block-adaptation

We now derive a version of the PDBS (14) adapted to the strucure of

$$F(x) = \sum_{j=1}^{m} F_j(x_j) \quad \text{and} \quad G_*(y) = \sum_{\ell=}^{n} G_{\ell *}(y_\ell),$$

where $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_n)$ in the (for simplicity) Hilbert spaces $X = \prod_{j=1}^{m} X_j$ and $Y = \prod_{\ell=1}^{n} Y_k$, and $F_j : X_j \to \overline{\mathbb{R}}$ and $G_{\ell *} : Y_\ell \to \overline{\mathbb{R}}$ are convex, proper, and lower semicontinuous.

For some "blockwise" step length parameters $\tau_j, \sigma_\ell > 0$ we take

$$J_X(x) = \sum_{j=1}^{m} \tau_j^{-1} N_{X_j}(x_j) \quad \text{and} \quad J_Y(y) = \sum_{\ell=1}^{n} \sigma_\ell^{-1} N_{Y_\ell}(y_\ell)$$

If K is now affine in y, (14) readily transforms into:

Block-adapted PDPS for K affine in y

Iteratively over $k \in \mathbb{N}$, for all $j = 1, \ldots, m$ and $\ell = 1, \ldots, m$, update:

$$\begin{aligned} x_j^{k+1} &:= \operatorname{prox}_{\tau_j F_j}(x_j^k - \tau_j \nabla_{x_j} K(x^k, y^k)), \\ y_\ell^{k+1} &:= \operatorname{prox}_{\sigma_\ell G_{\ell *}}(y_\ell^k + \sigma_\ell [2\nabla_{y_\ell} K(x^{k+1}, y^k) - \nabla_{y_\ell} K(x^k, y^k)]). \end{aligned} \tag{16}$$

The idea is that the blockwise step length parameters adapt the algorithm to the structure of the problem. We will return their choices in the examples of Section 4.3.

Remark 2 For bilinear K, (16) is the "diagonally preconditioned" method of [13], or an unaccelerated non-stochastic variant of the methods in [40]. For K affine in y, (16) differs from the methods in [32] by placing the over-relaxation in the dual step outside K, compare Remark 1.

Recall the saddle-point formulation (3) for inverse problems with nonlinear forward operators. We can now adapt step lengths to the constituent dual blocks:

Example 2 Let $A_1 \in C^1(X; Y_1^*)$ and $A_2 \in \mathbb{L}(X; Y_2^*)$, and suppose the convex functions $G_1 : Y_1^* \to \overline{\mathbb{R}}$ and $G_2 : Y_2^* \to \overline{\mathbb{R}}$ have the preconjugates G_{1*} and G_{2*}. Then we can write the problem

$$\min_{x \in X} \; G_1(A_1(x)) + G_2(A_2 x) + F(x).$$

in the form (S) with $G_*(y_1, y_2) = G_{1*}(y_1) + G_{2*}(y_2)$ and $K(x, y) = \langle A_1(x) | y_1 \rangle + \langle A_2 x | y_2 \rangle$. The algorithm (16) specialises as

$$\begin{aligned}x^{k+1} &:= \operatorname{prox}_{\tau F}(x^k - \tau[\nabla A_1(x^k)^* y_1 + A_2^* y_2]),\\ y_1^{k+1} &:= \operatorname{prox}_{\sigma_1 G_{1*}}(y_1^k + \sigma_1[2A_1(x^{k+1}) - A_1(x^k)]),\\ y_2^{k+1} &:= \operatorname{prox}_{\sigma_2 G_{2*}}(y_2^k + \sigma_2[A_2(2x^{k+1} - x^k)])\end{aligned}$$

for some step length parameters $\tau, \sigma_1, \sigma_2 > 0$ that we discuss further in Example 7. We also return to the local neighbourhood of convergence in Example 16.

4 Convergence theory

We now seek to understand when the basic version (13) of the PDBS convergences.

4.1 A fundamental estimate

We start with a simple estimate applicable to general methods of the form

$$0 \in H(u^{k+1}) + D_1 B(u^{k+1}, u^k) \tag{BP}$$

for some set-valued $H : U \rightrightarrows U^*$ and a Bregman divergence $B := B_J$ generated by some Gâteaux-differentiable $J : U \to \mathbb{R}$. We analyse (BP) following the "testing" ideas introduced in [39], extending them to the Bregman–Banach space setting, however in a simplified constant-metric setting that cannot model accelerated methods.

Theorem 1 *On a Banach space U, let $H : U \rightrightarrows U^*$, and let $B := B_J$ be generated by a Gâteaux-differentiable $J : U \to \mathbb{R}$. Suppose (BP) is solvable for $\{u^{k+1}\}_{k\in\mathbb{N}}$ given an initial iterate $u^0 \in U$. If for all $k = 0, \ldots, N-1$, for some $\bar{u} \in U$ and $\mathcal{G}(u^{k+1}, \bar{u}) \in \mathbb{R}$ the* fundamental condition

$$\langle h^{k+1} | u^{k+1} - \bar{u} \rangle \geq \mathcal{G}(u^{k+1}, \bar{u}) \quad (h^{k+1} \in H(u^{k+1})) \tag{C}$$

holds, then so do the quantitative Δ-Féjer monotonicity

$$B(\bar{u}, u^{k+1}) + B(u^{k+1}, u^k) + \mathcal{G}(u^{k+1}, \bar{u}) \leq B(\bar{u}, u^k) \tag{F}$$

and the descent inequality

$$B(\bar{u}, u^N) + \sum_{k=0}^{N-1} B(u^{k+1}, u^k) + \sum_{k=0}^{N-1} \mathcal{G}(u^{k+1}, \bar{u}) \leq B(\bar{u}, u^0). \tag{D}$$

The *generic gap functional* $\mathcal{G}(u^{k+1}, \bar{u})$ models any function value differences available from H. For example, if $H = \partial F$ for some convex function F, then (C) holds for $\mathcal{G}(u^{k+1}, \bar{u}) = F(u^{k+1}) - F(\bar{u})$. If our problem is non-convex, we will try to locally take $\mathcal{G}(u^{k+1}, \bar{u}) \geq 0$, so that (C) becomes a nonsmooth second-order growth condition.

Proof We can write (BP) as

$$0 = h^{k+1} + D_1 B(u^{k+1}, u^k) \quad \text{for some} \quad h^{k+1} \in H(u^{k+1}). \tag{17}$$

Testing (17) by applying $\langle \cdot | u^{k+1} - \bar{u} \rangle$ we obtain

$$0 = \langle h^{k+1} + D_1 B(u^{k+1}, u^k) | u^{k+1} - \bar{u} \rangle.$$

We use the three-point identity (6) to transform this into

$$B(\bar{u}, u^k) = \langle h^{k+1} | u^{k+1} - \bar{u} \rangle + B(\bar{u}, u^{k+1}) + B(u^{k+1}, u^k).$$

Inserting (C), we obtain (F). Summing the latter over $k = 0, \ldots, N-1$ yields (D). □

4.2 Ellipticity of the Bregman divergences

Besides (C), for Theorem 1 to we useful to prove the convergence of the PDBS, we need at least the semi-ellipticity of $B = B^0$. This is the present topic.

Standing assumption

In this subsection, we assume that B_X is τ^{-1}-elliptic and B_Y is σ^{-1}-elliptic for some $\tau, \sigma > 0$. This is true for the Hilbert-space PDPS (15) where τ and σ are the primal and dual step lengths.

The examples that follow the next general lemma will provide improved estimates.

Lemma 2 *Suppose $K \in C^1(X \times Y)$ is Lipschitz-continuously differentiable with the factor L_{DK} in a convex subdomain $\Omega \subset X \times Y$. Then for $u, u' \in \Omega$,*

$$B_K(u', u) \leq \frac{L_{DK}}{2} \|u' - u\|^2_{X \times Y}. \tag{18}$$

Consequently, if B_X is τ^{-1}-elliptic and B_Y is σ^{-1}-elliptic and $1 \geq \max\{\tau, \sigma\} L_{DK}$, then B^0 is semi-elliptic (elliptic if the the inequality is strict) within Ω.

Proof By definition, $B_K(u', u) = K(u') - K(u) - \langle DK(u) | u' - u \rangle$. Using the mean value equality in $\mathbb{R}$ with the chain rule and the Cauchy–Schwarz inequality, we get

$$B_K(u', u) = \int_0^1 \langle DK(u + t(u' - u)) - DK(u) | u' - u \rangle \, dt \leq \int_0^1 t L_{DK} \|u' - u\|^2_{X \times Y} \, dt.$$

Calculating the last integral yields (18).

For the (semi-)ellipticity, we need $B^0(u, u') \geq \frac{\varepsilon}{2} \|u - u'\|^2_{X \times Y}$ for some $\varepsilon > 0$ ($\varepsilon = 0$) and all $u, u' \in \Omega$. Since B_X and B_Y are τ^{-1}- and σ^{-1}-elliptic, we have

$$
\begin{aligned}
B^0(u',u) &= B_X(x',x) + B_Y(y',y) - B_K(u',u) \\
&\geq \frac{1}{2\tau}\|x'-x\|_X^2 + \frac{1}{2\sigma}\|y'-y\|_Y^2 - B_K(u'u).
\end{aligned} \tag{19}
$$

Using (18), therefore $B^0(u',u) \geq \frac{\tau^{-1}-L_{DK}}{2}\|x'-x\|_X^2 + \frac{\sigma^{-1}-L_{DK}}{2}\|y'-y\|_Y^2$. Thus B^0 is ε-elliptic when $\tau^{-1},\sigma^{-1} \geq L_{DK}+\varepsilon$. This gives the claim. ◻

Example 3 Suppose $K(x,y) = E(x)$ with DE L_{DE}-Lipschitz in $\Omega = X \times Y$. Then $L_{DK} = L_{DE}$, so we recover the standard-for-gradient-descent step length bound $1 \geq \tau L_{DE}$ for B^0 to be semi-elliptic in Ω (elliptic if the inequality is strict).

Example 4 If $K(x,y) = \langle Ax|y\rangle$ for $A \in \mathbb{L}(X;Y^*)$, then $L_{DK} = \|A\|$ within $\Omega = X\times Y$. In fact

$$
\langle DK(u+t(u'-u)) - DK(u)|u'-u\rangle = 2t\langle A(x-x')|y-y'\rangle.
$$

Therefore, taking any $w > 1$, we easily improve (18) to

$$
\begin{aligned}
B_K(u',u) &\leq \|A\|\|x'-x\|_X\|y'-y\|_Y \\
&\leq \frac{w\|A\|}{2}\|x'-x\|_X^2 + \frac{w^{-1}\|A\|}{2}\|y'-y\|_Y^2 \quad (u,u' \in X\times Y).
\end{aligned} \tag{20}
$$

By (19), B^0 is therefore ε-elliptic if $\tau^{-1} \geq w\|A\|+\varepsilon$ and $\sigma^{-1} \geq w^{-1}\|A\|+\varepsilon$. Taking $w = \sigma\|A\|/(1-\sigma\varepsilon)$ this holds if $1 \geq \tau\sigma\|A\|^2/(1-\sigma\varepsilon)+\tau\varepsilon$. Hence B^0 is elliptic under the standard-for-PDPS [12] step length condition $1 > \tau\sigma\|A\|^2$.

Example 5 Suppose $K(x,y) = \langle A(x)|y\rangle$ with A and DA Lipschitz with the respective factors $L_A, L_{DA} \geq 0$. Then, for any $w > 1$, using the mean value equality as in the proof of Lemma 2, we deduce

$$
\begin{aligned}
B_K(u',u) &= \langle A(x') - A(x)|y'\rangle - \langle DA(x)(x'-x)|y\rangle \\
&= \langle A(x') - A(x)|y'-y\rangle + \langle A(x') - A(x) - DA(x)(x'-x)|y\rangle \\
&\leq L_A\|x'-x\|_X\|y'-y\|_Y + \tfrac{L_{DA}\|y'\|}{2}\|x'-x\|_X^2 \\
&\leq \frac{wL_A + L_{DA}\|y\|}{2}\|x'-x\|_X^2 + \frac{w^{-1}L_A}{2}\|y'-y\|_Y^2.
\end{aligned} \tag{21}
$$

If $\rho_y > 0$ is such that $\|y\| \leq \rho_y$, taking $w = \sigma L_A/(1-\sigma\varepsilon)$, similarly to Example 4 we deduce B^0 to be elliptic within $\Omega = X \times B(0,\rho_y)$ if $1 > \tau\sigma L_A^2 + \tau\frac{L_{DA}\rho_y}{2}$.

We can combine the examples above:

Example 6 As in Example 2, take $K(x,(y_1,y_2)) = \langle A_1(x)|y_1\rangle + \langle A_2x|y_2\rangle$ with $A_1 \in C^1(X;Y_1^*)$ and $A_2 \in \mathbb{L}(X;Y_2^*)$. We bound B_K by summing (20) for A_1 and (21) for A_2. This yields for any $w_1, w_2 > 0$ the estimate

$$B_K(u',u) \le \frac{w_1 L_{A_1} + L_{DA_1}\|y_1\|}{2}\|x - x'\|_X^2 + \frac{w_1^{-1} L_{A_1}}{2}\|y_1' - y_1\|_Y^2 \qquad (22)$$
$$+ \frac{w_2\|A_2\|}{2}\|x' - x\|_X^2 + \frac{w_2^{-1}\|A_2\|}{2}\|y_2' - y_2\|_{Y_2}^2.$$

Taking $w_1 = \sigma L_{A_1}/(1-\sigma\varepsilon)$ and $w_2 = \sigma\|A_2\|/(1-\sigma\varepsilon)$, and using (19), we deduce that B^0 is elliptic (some $\varepsilon > 0$) of B^0 within Ω if $1 > \tau\sigma(L_{A_1}^2 + \|A_2\|^2) + \tau\frac{L_{DA_1}\rho_{y_1}}{2}$.

Remark 3 In Examples 5 and 6 we needed a bound on the dual variable y. In the latter, as an improvement, this was only needed on the subspace Y_1 of non-bilinearity. An ad-hoc solution is to introduce the bound into the problem. In the Hilbert case, [17, 18] secure such bounds by taking the primal step length τ small enough and arguing as in Theorem 1 individually on the primal and dual iterates.

4.3 Ellipticity for block-adapted methods

We now study ellipticity for block-adapted methods. The goal is to obtain faster convergence by adapting the blockwise step length parameters to the problem structure (connections between blocks) and the local (blockwise) properties of the problem.

Standing assumption

In this subsection, we assume F, G_*, J_X and J_Y to have the form of Section 3.3. In particular, X and Y are (products of) Hilbert spaces, and

$$B^0(u',u) = \sum_{j=1}^{m} \frac{1}{2\tau_j}\|x_j' - x_j\|_{X_j}^2 + \sum_{\ell=1}^{n} \frac{1}{2\sigma_\ell}\|y_\ell' - y_\ell\|_{Y_\ell}^2 - B_K(u',u). \qquad (23)$$

We start by refining the two-block Example 6 to be adapted to the blocks:

Example 7 Let $K(x,(y_1,y_2)) = \langle A_1(x)|y_1\rangle + \langle A_2 x|y_2\rangle$ with $A_1 \in C^1(X;Y_1^*)$ and $A_2 \in \mathbb{L}(X;Y_2^*)$ as in Examples 2 and 6. Write $\tau = \tau_1$. Using (22) in (23) for $m = 1$ and $n = 2$ with (22), we see B^0 to be ε-elliptic within $\Omega = X \times B(0,\rho_{y_1}) \times Y_2$ if $\tau^{-1} \ge w_1 L_{A_1} + L_{DA_1}\rho_{y_1} + w_2\|A_2\| + \varepsilon$ and $\sigma_1^{-1} \ge w_1^{-1} L_{A_1}$ as well as $\sigma_2^{-1} \ge w_2^{-1}\|A_2\| + \varepsilon$. Taking $w_1 = \sigma_1 L_{A_1}/(1-\sigma_1\varepsilon)$ and $w_2 = \sigma_2\|A_2\|/(1-\sigma_2\varepsilon)$, B^0 is therefore elliptic (some $\varepsilon > 0$) within Ω if $1 > \tau(\sigma_1 L_{A_1}^2 + \sigma_2\|A_2\|^2) + \tau\frac{L_{DA_1}\rho_{y_1}}{2}$.

Example 8 In Example 7, if both $A_1 \in \mathbb{L}(X;Y_1^*)$ and $A_2 \in \mathbb{L}(X;Y_2^*)$, then B^0 is elliptic within $\Omega = X \times Y_1 \times Y_2$ if $1 > \tau(\sigma_1\|A_1\|^2 + \sigma_2\|A_2\|^2)$.

Example 9 Suppose we can write $K(x,y) = \sum_{j=1}^m \sum_{\ell=1}^n K_{j\ell}(x_j, y_\ell)$ with each $K_{j\ell}$ Lipschitz-continuously differentiable with the factor $L_{j\ell}$. Following Lemma 2,

$$B_K(u',u) \le \sum_{j=1}^{m}\sum_{\ell=1}^{n} \frac{L_{j\ell}}{2}(\|x'_j - x_j\|^2 + \|y'_\ell + y_\ell\|^2). \tag{24}$$

Consequently, using (23), we see that B^0 is ε-elliptic if $1 \ge \tau_j(\sum_{\ell=1}^n L_{j\ell} + \varepsilon)$ and $1 \ge \sigma_\ell(\sum_{j=1}^n L_{j\ell} + \varepsilon)$ for all $j = 1, \ldots, m$ and $\ell = 1, \ldots, n$.

Example 10 If $K(x,y) = \sum_{j=1}^m \sum_{\ell=1}^m \langle A_{j\ell} x_j | y_\ell \rangle$ for some $A_{j\ell} \in \mathbb{L}(X_j; Y_\ell^*)$, then following Example 4, for arbitrary $w_{j\ell} > 0$,

$$\begin{aligned} B_K(u',u) &\le \sum_{j=1}^{m}\sum_{\ell=1}^{m} \|A_{j\ell}\|\|x'_j - x_j\|\|y'_j - y_j\| \\ &\le \sum_{j=1}^{m}\sum_{\ell=1}^{n} \left(\frac{w_{j\ell}\|A_{j\ell}\|}{2}\|x'_j - x_j\|^2 + \frac{w_{j\ell}^{-1}\|A_{j\ell}\|}{2}\|y'_\ell - x_\ell\|^2 \right). \end{aligned}$$

Using (23), B^0 is thus ε-elliptic if $1 \ge \tau_j(\varepsilon + \sum_{\ell=1}^n w_{j\ell}\|A_{j\ell}\|)$ and $1 \ge \sigma_\ell(\varepsilon + \sum_{j=1}^m w_{j\ell}^{-1}\|A_{j\ell}\|)$ for all $j = 1, \ldots, m$ and $\ell = 1, \ldots, n$. We can use the factors $w_{j\ell}$ to adapt the algorithm to the different blocks for potentially better convergence.

4.4 Non-smooth second-order conditions

We now study conditions for (C) to hold with $\mathcal{G}(\,\cdot\,, \bar u) \ge 0$. We start by writing out the condition for the PDBS.

Lemma 3 *Let $\bar u \in X \times Y$ and suppose for some $\mathcal{G}(u,\bar u) \in \mathbb{R}$ and a neighbourhood $\Omega_{\bar u} \subset X \times Y$ that for all $u \in \Omega_{\bar u}$, $x^* \in \partial F(x)$, and $y^* \in \partial G_*(y)$,*

$$\langle x^* + D_x K(x,y) | x - \bar x\rangle + \langle y^* - D_y K(x,y) | y - \bar y\rangle \ge \mathcal{G}(u,\bar u). \tag{C^2}$$

Let $\{u^{k+1}\}_{k\in\mathbb{N}}$ be generated by the PDBS (14) for some $u^0 \in X \times Y$, and suppose $\{u^k\}_{k\in\mathbb{N}} \subset \Omega_{\bar u}$. Then the fundamental condition (C), the quantitative Δ-Féjer monotonicity (F), and the descent inequality (D) hold.

Proof Theorem 1 proves (F) and (D) if we show (C^2). For H in (10), we have

$$h^{k+1} = \begin{pmatrix} x^*_{k+1} + D_x K(x^{k+1}, y^{k+1}) \\ y^*_{k+1} - D_y K(x^{k+1}, y^{k+1}) \end{pmatrix} \in H(u^{k+1}) \quad \text{with} \quad \begin{cases} x^*_{k+1} \in \partial F(x^{k+1}), \\ y^*_{k+1} \in \partial G_*(y^{k+1}). \end{cases}$$

Thus (C) expands as (C^2) for $u = u^{k+1}$ and $(x^*, y^*) = (x^*_{k+1}, y^*_{k+1})$. □

In Section 4.7 on the convergence of gap functionals, we will consider general $\bar u$ in (C^2). For the moment, we however fix a root $\bar u = \hat u \in H^{-1}(0)$. Then

$$0 = \begin{pmatrix} \hat x^* + D_x K(\hat x, \hat y) \\ \hat y^* - D_y K(\hat x, \hat y) \end{pmatrix} \in H(\hat u) \quad \text{with} \quad \begin{cases} \hat x^* \in \partial F(\hat x), \\ \hat y^* \in \partial G_*(\hat y). \end{cases} \tag{25}$$

Since we assume F and G_* to be convex, their subdifferentials are monotone. When K is not convex-concave, and to obtain strong convergence of iterates even when it is, we will need some strong monotonicity of the subdifferentials, but only *at* a solution. Specifically, for $\gamma > 0$, we say that $T : X \rightrightarrows X^*$ is *γ-strongly monotone at* $\hat{x}$ for $\hat{x}^* \in T(\hat{x})$ if

$$\langle x^* - \hat{x}^* | x - \hat{x} \rangle \geq \gamma \|x - \hat{x}\|_X^2 \quad (x \in X,\ x^* \in T(x)). \tag{26}$$

If $\gamma = 0$, we drop the word "strong". For $T = \partial F$, (26) follows from the γ-strong subdifferentiability of F.

Standing assumption

Throughout the rest of this subsection, we assume (25) to hold and that ∂F is (γ_F-strongly) monotone at $\hat{x}$ for $\hat{x}^*$, and ∂G_* is (γ_{G_*}-strongly) monotone at $\hat{y}$ for $\hat{y}^*$.

Lemma 4 *The nonsmooth second-order growth condition* (C^2) *holds provided*

$$\gamma_F \|x - \hat{x}\|^2 + \gamma_{G_*} \|y - \hat{y}\|^2 \geq B_K(\hat{u}, u) + B_K(u, \hat{u}) + \mathcal{G}(u, \hat{u}) \quad (u \in \Omega_{\bar{u}}), \tag{27}$$

equivalently

$$\gamma_F \|x - \hat{x}\|^2 + \gamma_{G_*} \|y - \hat{y}\|^2 \geq d_K(\hat{u}, u) + d_K(u, \hat{u}) + \mathcal{G}(u, \hat{u}) \quad (u \in \Omega_{\bar{u}}) \tag{27'}$$

for

$$d_K(u, \bar{u}) := K(x, y) - K(\bar{x}, \bar{y}) + \langle D_x K(x, y) | \bar{x} - x \rangle + \langle D_y K(\bar{x}, \bar{y}) | \bar{y} - y \rangle. \tag{28}$$

Note that (27) involves the *symmetrised Bregman divergence* $B_K^S(u, u') := B_K(u, u') + B_K(u', u)$ generated by K.

Proof Inserting the zero of (25) in (C^2), we rewrite the latter as

$$\begin{aligned}\langle x^* - \hat{x}^* | x - \hat{x} \rangle + \langle y^* - \hat{y}^* | y - \hat{y} \rangle &\geq \langle D_x K(x, y) - D_x K(\hat{x}, \hat{y}) | \hat{x} - x \rangle \\ &\quad + \langle D_y K(x, y) - D_y K(\hat{x}, \hat{y}) | y - \hat{y} \rangle + \mathcal{G}(u^{k+1}, \hat{u}).\end{aligned}$$

Using the assumed strong monotonicities, and the definitions of B_K and d_K, this is immediately seen to hold when (27) or (27′) does. ◻

Example 11 If K is convex-concave, the next Lemma 5 and Lemma 4 prove (C^2) for

$$\mathcal{G}(u, \hat{u}) = \gamma_F \|x - \hat{x}\|^2 + \gamma_{G_*} \|y - \hat{y}\|^2 \geq 0 \quad \text{within} \quad \Omega_{\hat{u}} = X \times Y.$$

This is in particular true for $K(x, y) = \langle Ax | y \rangle + E(x)$ with $A \in \mathbb{L}(X; Y^*)$ and $E \in C^1(X)$ convex.

Lemma 5 *Suppose $K : X \times Y \to \mathbb{R}$ is Gâteaux-differentiable and convex-concave. Then $d_K(u, \bar{u}) \leq 0$ and $B_K^S(u, \bar{u}) \leq 0$ for all $u, \bar{u} \in X \times Y$.*

Proof The convexity of $K(\cdot, y)$ and the concavity of $K(\bar{x}, \cdot)$ show

$$K(x,y) - K(\bar{x},y) + \langle D_x K(x,y)|\bar{x}-x\rangle \le 0 \quad \text{and}$$
$$K(\bar{x},y) - K(\bar{x},\bar{y}) + \langle D_y K(\bar{x},\bar{y})|\bar{y}-y\rangle \le 0.$$

Summing these two estimates proves $d_K(u,\bar{u}) \le 0$, consequently $B^S_K(u,\bar{u}) = d_K(u,\bar{u}) + d_K(\bar{u},u) \le 0$. □

Example 12 Supppose K has L_{DK}-Lipschitz derivative within $\Omega \subset X \times Y$. If $\hat{u} \in \Omega$, then by Lemma 2, $B_K(u,\hat{u}), B_K(\hat{u},u) \le \frac{L_{DK}}{2}\|u-\hat{u}\|^2_{X\times Y}$ for $u \in \Omega$. Thus (C^2) holds by Lemma 4 with $\Omega_{\hat{u}} = \Omega$ and

$$\mathcal{G}(u,\hat{u}) = (\gamma_F - L_{DK})\|x-\hat{x}\|^2 + (\gamma_{G_*} - L_{DK})\|y-\hat{y}\|^2.$$

This is non-negative if $\gamma_F, \gamma_{G_*} \ge L_{DK}$.

Example 13 Let $K(x,y) = \langle A(x)|y\rangle$ for some $A \in \mathbb{L}(X;Y^*)$ such that DA is Lipschitz with the factor $L_{DA} \ge 0$. For some $\tilde{\gamma}_F, \tilde{\gamma}_{G_*} \ge 0$ and $\rho_y, \hat{\rho}_x, \alpha > 0$, let either

(a) $\tilde{\gamma}_F \ge \frac{L_{DA}}{2}(\rho_y + \|\hat{y}\|_Y)$, $\tilde{\gamma}_{G_*} \ge 0$, and $\Omega_{\hat{u}} = X \times B(0,\rho_y)$; or
(b) $\tilde{\gamma}_F > L_{DA}\left(\|\hat{y}\|_Y + \frac{\alpha}{2}\right)$, $\tilde{\gamma}_{G_*} \ge \frac{L_{DA}}{2\alpha}\hat{\rho}_x^2$, and $\Omega_{\hat{u}} = B(\hat{x},\hat{\rho}_x) \times Y$.

Then Lemma 4 proves (C^2) with

$$\mathcal{G}(u,\hat{u}) = (\gamma_F - \tilde{\gamma}_F)\|x-\hat{x}\|^2 + (\gamma_{G_*} - \tilde{\gamma}_{G_*})\|y-\hat{y}\|^2.$$

To see this, we need to prove (27′). Now

$$d_K(u,\hat{u}) := \langle A(x) - A(\hat{x}) + DA(x)(\hat{x}-x)|y\rangle \quad (u,\hat{u} \in X \times Y). \tag{29}$$

Arguing with the mean value equality and the Lipschitz assumption as in Lemma 2, we get $d_K(\hat{u},u) + d_K(u,\hat{u}) \le \frac{L_{DA}}{2}(\|y\|_Y + \|\hat{y}\|_Y)\|x-\hat{x}\|^2$. Thus (a) implies (27′). By (29), the mean-value equality, and the Lipschitz assumption, also

$$\begin{aligned} d_K(u,\hat{u}) + d_K(\hat{u},u) &= \langle [DA(x) - DA(\hat{x})](\hat{x}-x)|\hat{y}\rangle \\ &\quad + \langle A(x) - A(\hat{x}) + DA(x)(\hat{x}-x)|y-\hat{y}\rangle \\ &\le L_{DA}\|x-\hat{x}\|^2_X\left(\|\hat{y}\|_Y + \tfrac{1}{2}\|y-\hat{y}\|_Y\right). \end{aligned}$$

Using Cauchy's inequality and (b) we deduce (27′).

Remark 4 In the last two examples, we need to bound some of the iterates, and to initialise close enough to a solution. Showing that the iterates stay in a local neighbourhood is a large part of the work in [17, 18], as discussed in Remark 3.

4.5 Second-order growth conditions for block-adapted methods

We now study second-order growth for problems with block structure.

Standing assumption

In this subsection, F and G_* are as in Section 3.3, each component subdifferential ∂F_j now (γ_{F_j}-strongly) monotone at $\hat{x}_j$ for $\hat{x}^*_j$ and each $\partial G_{\ell *}$ ($\gamma_{G_{\ell *}}$-strongly) monotone at $\hat{y}_\ell$ for $\hat{y}^*_\ell$. Here $\hat{x}_j$, $\hat{x}^*_j$, $\hat{y}_\ell$ and $\hat{y}^*_\ell$ are the components of $\hat{x}$, $\hat{x}^*$, $\hat{y}$, and $\hat{y}^*$ in the corresponding subspace, assumed to satisfy the critical point condition (25).

As only some of the component functions may have $\gamma_{F_j}, \gamma_{G_{\ell *}} > 0$, through detailed analysis of the block structure, we hope to obtain (strong) convergence on some subspaces even if the entire primal or dual variables might not converge.

Similarly to Lemma 4 we prove:

Lemma 6 *Suppose for some neighbourhood $\Omega_{\hat{u}} \subset X \times Y$ that*

$$\Delta_{k+1} := \sum_{j=1}^{m} \tilde{\gamma}_{F_j} \|x_j - \hat{x}_j\|^2_{X_j} + \sum_{\ell=1}^{n} \tilde{\gamma}_{G_{\ell *}} \|y_\ell - \hat{y}_\ell\|^2_{Y_\ell} \geq d_K(\hat{u}, u) + d_K(u, \hat{u})$$

for some $\tilde{\gamma}_{F_j}, \gamma_{G_{\ell *}} \geq 0$ for all $u \in \Omega_{\hat{u}}$. Then (C^2) holds with

$$\mathcal{G}(u, \hat{u}) = \sum_{j=1}^{m} (\gamma_{F_j} - \tilde{\gamma}_{F_j}) \|x_j - \hat{x}_j\|^2_{X_j} + \sum_{\ell=1}^{n} (\gamma_{G_{\ell *}} - \tilde{\gamma}_{G_{\ell *}}) \|y_\ell - \hat{y}_\ell\|^2_{Y_\ell}. \tag{30}$$

In the convex–concave case, we can transfer all strong monotonicity into $\mathcal{G}$:

Example 14 If K is convex-concave, then by Lemmas 5 and 6, (C^2) holds with $\Omega_{\hat{u}} = X \times Y$ and $\mathcal{G}$ as in (30) for $\tilde{\gamma}_{F_j} = 0$ and $\tilde{\gamma}_{G_{\ell *}} = 0$. We have $\mathcal{G}(\,\cdot\,, \hat{u}) \geq 0$ always.

Example 15 As in Example 9, suppose we can write $K(x, y) = \sum_{j=1}^{m} \sum_{\ell=1}^{n} K_{j\ell}(x_j, y_\ell)$ with each $K_{j\ell}$ Lipschitz-continuously differentiable with the factor $L_{j\ell}$ in Ω. Then using (24) and Lemma 6, we see (C^2) to hold with $\Omega_{\hat{u}} = \Omega$ and $\mathcal{G}$ as in (30) with

$$\tilde{\gamma}_{F_j} = \sum_{\ell=1}^{n} L_{j\ell} \quad (j = 1, \ldots, m) \quad \text{and} \quad \tilde{\gamma}_{G_{\ell *}} = \sum_{j=1}^{m} L_{j\ell} \quad (\ell = 1, \ldots, n).$$

Thus $\mathcal{G}(\,\cdot\,, \hat{u}) \geq 0$ if $\gamma_{F_j} \geq \sum_{\ell=1}^{n} L_{j\ell}$ and $\gamma_{G_{\ell *}} \geq \sum_{j=1}^{m} L_{j\ell}$ for all ℓ and j.

The special case of Example 9 with each $K_{j\ell}$ bilinear, corresponding to Example 10 for ellipticity, is covered by Example 14.

We consider in detail the two dual block setup of Examples 2 and 7:

Example 16 As in Example 2, let $K(x, y) = \langle A_1(x) | y_1 \rangle + \langle A_2 x | y_2 \rangle$ for $A_1 \in C^1(X; Y_1^*)$ and $A_2 \in \mathbb{L}(X; Y_2^*)$. Then, as in (29),

$$d_K(u, \bar{u}) = \langle A_1(x) - A_1(\bar{x}) + DA_1(x)(\bar{x} - x) | y_1 \rangle,$$

which does not depend on A_2. For any $\alpha, \rho_y, \hat{\rho}_x > 0$ let either

(a) $\tilde\gamma_F \ge \frac{L_{DA_1}}{2}(\rho_{y_1} + \|\hat y_1\|_{Y_1})$, $\tilde\gamma_{G_{1*}} \ge 0$, and $\Omega_{\hat u} = X \times B(0, \rho_{y_1})$; or
(b) $\tilde\gamma_F > L_{DA_1}\left(\|\hat y_1\|_{Y_1} + \frac{\alpha}{2}\right)$, $\tilde\gamma_{G_{1*}} \ge \frac{L_{DA_1}}{2\alpha}\hat\rho_x^2$, and $\Omega_{\hat u} = B(\hat x, \hat\rho_x) \times Y$.

Arguing as in Example 13 and using Lemma 6, we then see (C^2) to hold with $\mathcal{G}$ as in (30) and $\tilde\gamma_{G_{2*}} = 0$. In this case $\mathcal{G}(\,\cdot\,, \hat u)$ is non-negative if $\gamma_F \ge \tilde\gamma_F$ and $\gamma_{G_{1*}} \ge \tilde\gamma_{G_{1*}}$.

4.6 Convergence of iterates

We are now ready to prove the convergence of the iterates. We start with weak convergence and proceed to strong and linear convergence. For weak convergence in infinite dimensions, we need some further technical assumptions. We recall that a set-valued map $T : X \rightrightarrows X^*$ is weak-to-strong (weak-$*$-to-strong) outer semicontinuous if $x_k^* \in T(x^k)$ and $x^k \rightharpoonup x$ ($x^k \overset{*}{\rightharpoonup} x$) and $x_k^* \to x^*$ imply $x^* \in T(x)$.

Assumption 1 Each of the spaces X and Y is, individually, either a reflexive Banach space or the dual of separable space. The operator $H : X \times Y \rightrightarrows X^* \times Y^*$ is weak(-$*$)-to-strong outer semicontinuous, where we mean by "weak(-$*$)" that we take the weak topology if the space is reflexive and weak-$*$ otherwise, individually on X and Y.

Subdifferentials of lower semicontinuous convex functions are weak(-$*$)-to-strong outer semicontinuous[5], so the outer semicontinuity of H depends mainly on K.

Example 17 If X and Y are finite-dimensional, Assumption 1 holds if $K \in C^1(X; Y)$.

Example 18 More generally, Assumption 1 holds if $K \in C^1(X \times Y)$ and DK is continuous from the weak(-$*$) topology to the strong topology.

Example 19 If $K = \langle Ax|y\rangle + E(x)$ for $A \in \mathbb{L}(X; Y^*)$ and $E \in C^1(X)$ convex, then H satisfies Assumption 1. Indeed, it can be shown that H is maximal monotone, hence weak(-$*$) outer semicontinuous similarly to convex subdifferentials.

> Verification of the conditions

To verify the nonsmooth second-order growth condition (C^2) for each of the following Theorems 2 to 4, we point Sections 4.4 and 4.5. For the verification of the (semi-)ellipticity of B^0, we point to Sections 4.2 and 4.3. As special cases of the PDBS (14), the theorems apply to the Hilbert-space PDPS (15) and its block-adaptation (16). Then J_X and J_Y are continuously differentiable and convex.

---

[5] This result seems difficult to find in the literature for Banach spaces, but follows easily from the definition of the subdifferential: If $F(x) \ge F(x^k) + \langle x_k^* | x - x^k\rangle$ and $x_k^* \to \hat x^*$ as well as $x^k \rightharpoonup$ (or $\overset{*}{\rightharpoonup}$) $\hat x$, then, using the fact that $\{\|x^k - \hat x\|\}_{k\in\mathbb{N}}$ is bounded, in the limit $F(x) \ge F(\hat x) + \langle \hat x^* | x - \hat x\rangle$.

Theorem 2 (Weak convergence) *Let F and G_* be convex, proper, and lower semicontinuous; $K \in C^1(X \times Y)$; and both $J_X \in C^1(X)$ and $J_Y \in C^1(Y)$ convex. Suppose Assumption 1 holds and for some $\hat{u} \in H^{-1}(0)$ that*

(i) (C²) *holds with $\mathcal{G}(\,\cdot\,, \hat{u}) \geq 0$ within $\Omega_{\hat{u}} \subset X \times Y$; and*
(ii) B^0 is elliptic within $\Omega \ni \hat{u}$.

Let $\{u^{k+1}\}_{k\in\mathbb{N}}$ be generated by the PDBS (14) *for any initial u^0, and suppose $\{u^k\}_{k\in\mathbb{N}} \subset \Omega \cap \Omega_{\hat{u}}$. Then there exists at least one cluster point of $\{u^k\}_{k\in\mathbb{N}}$, and all weak(-∗) cluster points belong to $H^{-1}(0)$.*

Proof Lemma 3 establishes (D). With $\varepsilon > 0$ the factor of ellipticity of B^0, it follows

$$\frac{\varepsilon}{2}\|\hat{u} - u^N\|^2_{X\times Y} + \frac{\varepsilon}{2}\sum_{k=0}^{N-1}\|u^{k+1} - u^k\|^2_{X\times Y} \leq B^0(\hat{u}, u^0) \quad (N \in \mathbb{N}).$$

Clearly $\|u^{k+1} - u^k\| \to 0$ while $\{\|u^N - \hat{u}\|\}_{N\in\mathbb{N}}$ is bounded. Using the Eberlein–Šmulyan theorem in a reflexive X or Y, and the Banach–Alaoglu theorem otherwise (X or Y the dual of a separable space), we may therefore find a subsequence of $\{u^N\}_{N\in\mathbb{N}}$ converging weakly(-∗) to some $\bar{x}$. Since $J^0 \in C^1(X \times Y)$, we deduce $D_1 B^0(u^{k+1}, u^k) \to 0$. Consequently (13) implies that $0 \in \limsup_{k\to\infty} H(u^{k+1})$, where we write "lim sup" for the Painlevé–Kuratowski outer limit of a sequence of sets in the strong topology. Since H is weak(-∗)-to-strong outer semicontinuous by Assumption 1, it follows that $0 \in H(\hat{u})$. Therefore, there exists at least one cluster point of $\{u^k\}_{k\in\mathbb{N}}$ belonging to $H^{-1}(0)$. Repeating the argument on any weak(-∗) convergent subsequence, we deduce that all cluster points belong to $H^{-1}(0)$. □

Remark 5 For a unique weak limit we may in Hilbert spaces use the quantitative Féjer monotonicity (F) with Opial's lemma [34, 9]. For bilinear K the result is relatively immediate, as B^0 is a squared matrix-weighted norm; see [39]. Otherwise a variable-metric Opial's lemma [17] and additional work based on the Brezis–Crandall–Pazy lemma [8, Corollary 20.59 (iii)] is required; see [17] for $K(x, y) = \langle A(x)|y\rangle$, and [18] for general K.

Theorem 3 (Strong convergence) *Let F and G_* be convex, proper, and lower semicontinuous; $K \in C^1(X \times Y)$; and both $J_X \in C(X)$ and $J_Y \in C(Y)$ convex and Gâteaux-differentiable. Suppose for some $\hat{u} \in H^{-1}(0)$ that*

(i) (C²) *holds with $\mathcal{G}(\,\cdot\,, \hat{u}) \geq 0$ within $\Omega_{\hat{u}} \subset X \times Y$; and*
(ii) B^0 is semi-elliptic within $\Omega \ni \hat{u}$.

Let $\{u^{k+1}\}_{k\in\mathbb{N}}$ be generated by the PDBS (14) *for any initial u^0, and suppose $\{u^k\}_{k\in\mathbb{N}} \subset \Omega \cap \Omega_{\hat{u}}$. Then $\mathcal{G}(u^{k+1}, \hat{u}) \to 0$ as $N \to \infty$.*

In particular, if $\mathcal{G}(u, \hat{u}) \geq \|P(u - \hat{u})\|^2_Z$ for some $P \in \mathbb{L}(X; Z)$, then $Px^N \to P\hat{x}$ strongly in Z and the ergodic sequence $\tilde{x}^N_P := \frac{1}{N}\sum_{k=0}^{N-1} Px^{k+1} \to P\hat{x}$ *at rate $O(1/N)$.*

Proof Lemma 3 establishes (D). By the semi-ellipticity of B^0 then $\sum_{k=0}^{N-1}\mathcal{G}(u^{k+1}, \hat{u}) \leq B^0(\hat{u}, u^0)$, $(N \in \mathbb{N})$. Since $\mathcal{G}(u^{k+1}, \hat{u}) \geq 0$, this shows that $\mathcal{G}(u^N, \hat{u}) \to 0$. The strong convergence of the primal variable for quadratically minorised $\mathcal{G}$ is then immediate whereas following by Jensen's inequality gives the ergodic convergence claim. ◻

Example 20 In Section 4.4, we can take $Pu = \sqrt{\gamma_F - \tilde\gamma_F}x$ if $\gamma_F > \tilde\gamma_F$ or $Pu = \sqrt{\gamma_{G_*} - \tilde\gamma_{G_*}}y$ if $\gamma_{G_*} > \tilde\gamma_{G_*}$. The examples of Section 4.5 for $x = (x_1, \ldots, x_m)$, $y = (y_1, \ldots, y_n)$ may allow $Pu = \sqrt{\gamma_{F_j} - \tilde\gamma_{F_j}}x_j$ or $Pu = \sqrt{\gamma_{G_{\ell *}} - \tilde\gamma_{G_{\ell *}}}y_\ell$.

Remark 6 Under similar conditions as Theorem 3, it is possible to obtain $O(1/N^2)$ convergence rates; see [12, 39] for the convex-concave case and [17, 18] in general.

Theorem 4 (Linear convergence) *Let F and G_* be convex, proper, and lower semicontinuous; $K \in C^1(X \times Y)$; and both $J_X \in C(X)$ and $J_Y \in C(Y)$ convex and Gâteaux-differentiable. Suppose for some $\gamma > 0$ and $\hat u \in H^{-1}(0)$ that*

(i) (C^2) holds with $\mathcal{G}(u, \hat u) \geq \gamma B^0(\hat u, u)$ within $\Omega_{\hat u} \subset X \times Y$; and
(ii) B^0 is elliptic within $\Omega \supset \hat u$.

Let $\{u^{k+1}\}_{k\in\mathbb{N}}$ be generated by the PDBS (14) for any initial u^0, and suppose $\{u^k\}_{k\in\mathbb{N}} \subset \Omega \cap \Omega_{\hat u}$. Then $B^0(\hat u, u^N) \to 0$ and $u^N \to \hat u$ at a linear rate.

In particular, if $\mathcal{G}(u, \hat u) \geq \gamma\|u - \hat u\|^2$, ($k \in \mathbb{N}$), for some $\gamma > 0$, and J^0 is Lipschitz-continuously differentiable, then $u^N \to \hat u$ at a linear rate.

Proof Lemma 3 establishes the quantitative Δ-Féjer monotonicity (F). Using (i), this yields $(1+\gamma)B^0(\hat u, u^{k+1}) \leq B^0(\hat u, u^k)$. By the semi-ellipticity of B^0, the claimed linear convergence of $B^0(\hat u, u^N) \to 0$ follows. Since B^0 is assumed elliptic, also $u^N \to \hat u$ linearly. If J^0 is Lipschitz-continuously differentiable, then, similarly to Lemma 2, $B^0(\hat u, u^{k+1}) \leq L_{DJ}\|u^{k+1} - \hat u\|^2$ for some $L_{DJ} > 0$. Thus $\mathcal{G}(u^{k+1}, \hat u) \geq \gamma L_{DJ}^{-1} B^0(\hat u, u^{k+1})$, so the main claim establishes the particular claim. ◻

Example 21 J^0 is Lipschitz-continuously differentiable if X and Y are Hilbert spaces with $J_X = \tau^{-1}N_X$ and $J_Y = \sigma^{-1}N_Y$, and K Lipschitz-continuously differentiable.

4.7 Convergence of gaps in the convex-concave setting

We finish this section by studying the convergence of gap functionals in the convex-concave setting.

Lemma 7 *Suppose F and G_* are convex, proper, and lower semicontinuous, and $K \in C^1(X \times Y)$ is convex-concave on $\operatorname{dom} F \times \operatorname{dom} G_*$. Then (C^2) holds for all $\bar u \in X \times Y$ with $\Omega_{\bar u} = X \times Y$ and $\mathcal{G} = \mathcal{G}^{\mathcal{L}}$ the* Lagrangian gap

$$\begin{aligned}\mathcal{G}^{\mathcal{L}}(u, \bar u) &:= \mathcal{L}(x, \bar y) - \mathcal{L}(\bar x, y) \\ &= [F(x) + K(x, \bar y) - G_*(\bar y)] - [F(\bar x) + K(\hat x, y) - G_*(y)].\end{aligned}$$

This functional is non-negative if $\bar u \in H^{-1}(0)$.

Moreover, if $\sum_{k=0}^{N-1} \mathcal{G}^{\mathcal{L}}(u^{k+1}, \bar u) \leq M(\bar u)$ for some $M(\bar u) \geq 0$, for all $\bar u \in X \times Y$ and all $N \in \mathbb{N}$, and we define the ergodic sequence $\tilde u^N := \frac{1}{N}\sum_{k=0}^{N-1} u^{k+1}$, *then*

(i) $0 \le \frac{1}{N}\sum_{k=0}^{N-1} \mathcal{G}^{\mathcal{L}}(u^{k+1}, \hat{u}) \to 0$ *at the rate* $O(1/N)$ *for* $\hat{u} \in H^{-1}(0)$.
(ii) $0 \le \mathcal{G}^{\mathcal{L}}(\tilde{u}^N, \hat{u}) \to 0$ *at the rate* $O(1/N)$ *for* $\hat{u} \in H^{-1}(0)$.
(iii) If $M \in C(X \times Y)$ *and* $\Omega \subset X \times Y$ *is bounded with* $\Omega \cap H^{-1}(0) \neq \emptyset$, *then* $0 \le \mathcal{G}_\Omega(\tilde{u}^N) \to 0$ *at the rate* $O(1/N)$ *for the* partial gap $\mathcal{G}_\Omega(u) := \sup_{\bar{u}\in\Omega} \mathcal{G}^{\mathcal{L}}(u, \bar{u})$.

The convergence results in Lemma 7 are *ergodic* because they apply to sequences of running averages. To understand the partial gap, we recall that with $K(x,y) = \langle Ax|y\rangle$ bilinear Fenchel–Rockafellar's theorem show that the *duality gap* $\mathcal{G}^D(u) := [F(x) + G_*(Ax)] + [F_*(-A^*y) + G_*^*(y)] \ge 0$ and is zero *if and only if* $u \in H^{-1}(0)$. The duality gap can be written $\mathcal{G}^D(u) = \mathcal{G}_{X\times Y}(u)$.

Proof By the convex-concavity of K and the definition of the subdifferential,

$$\langle D_x K(x,y)|x - \bar{x}\rangle - \langle D_y K(x,y)|y - \bar{y}\rangle \\ \ge [K(x,y) - K(\bar{x},y)] - [K(x,y) - K(x,\bar{y})] = K(x,\bar{y}) - K(\bar{x},y).$$

for all $(x,y) \in X \times Y$. Also using $x^* \in \partial F(x^{k+1})$ and $y^* \in \partial G_(y^{k+1})$ with the definition of the convex subdifferential, we see that $\mathcal{G} = \mathcal{G}^{\mathcal{L}}$ satisfies (C^2). The non-negativity of $\mathcal{G}(\,\cdot\,, \hat{u})$ follows by similar reasoning, first using that

$$K(x,\hat{y}) - K(\hat{x},y) \ge \langle D_x K(\hat{x},\hat{y})|x - \hat{x}\rangle - \langle D_y K(\hat{x},\hat{y})|y - \hat{y}\rangle \tag{31}$$

for all $(x,y) \in X \times Y$, and following by the definition of the subdifferential applied to $-D_x K(\hat{x},\hat{y}) \in \partial F(\hat{x})$ and $D_y K(\hat{x},\hat{y}) \in \partial G_*(\hat{y})$.

For (i)–(iii), we first observe that the semi-ellipticity of B^0 and (C^2) imply $\sum_{k=0}^{N-1} \mathcal{G}^{\mathcal{L}}(u^{k+1}, \bar{u}) \le M(\bar{u})$. Dividing by N and using that $\mathcal{G}^{\mathcal{L}}(u^{k+1}, \hat{u}) \ge 0$ for $\bar{u} \in H^{-1}(0)$, we obtain (i). Jensen's inequality then gives $\mathcal{G}^{\mathcal{L}}(\tilde{u}^{k+1}, \bar{u}) \le M(\bar{u})/N$, hence (ii) for $\bar{u} \in H^{-1}(0)$. Finally, taking the supremum over $\bar{u} \in \Omega$ gives (iii) because M is bounded on bounded sets. ◻

In the following theorem, we may in particular take $K(x,y) = \langle Ax|y\rangle$ bilinear, or $K(x,y) = \langle Ax|y\rangle + E(x)$ with E convex. Lemma 2 and Examples 3 and 4 provide step length conditions that ensure the semi-ellipticity required of B^0 in Theorem 5.

Theorem 5 (Gap convergence) *Let* $F : X \to \overline{\mathbb{R}}$ *and* $G_* : Y \to \overline{\mathbb{R}}$ *be convex, proper, and lower semicontinuous;* $K \in C^1(X \times Y)$ *convex-concave within* dom $F \times$ dom G_**;* $J_X \in C^1(X)$ *and* $J_Y \in C^1(Y)$ *convex. If* B^0 *is semi-elliptic, then the iterates* $\{u^{k+1}\}_{k\in\mathbb{N}}$ *generated by the PDBS* (14) *for any initial* $u^0 \in X \times Y$ *satisfy Lemma 7 (i)–(iii).*

Proof By Lemma 7, holds with $\mathcal{G} = \mathcal{G}^{\mathcal{L}}$ Hence by Lemma 3, (D) holds. Since B^0 is semi-elliptic, this implies that that $\sum_{k=0}^{N-1} \mathcal{G}(u^{k+1}, \bar{u}) \le M(\bar{u}) := B^0(\bar{u}, u^0)$ for all $N \in \mathbb{N}$. Since J_X, J_Y, and K are continuously differentiable, $M \in C^1(X \times Y)$. The rest follows from the second part of Lemma 7. ◻

5 Inertial terms

We now generalise (BP), making the involved Bregman divergences dependent on the iteration k and earlier iterates:

$$0 \in H(u^{k+1}) + D_1 B_{k+1}(u^{k+1}, u^k) + D_1 B^-_{k+1}(u^k, u^{k-1}), \tag{IPP}$$

for $B_{k+1} := B_{J_{k+1}}$ and $B^-_{k+1} := B_{J^-_{k+1}}$ generated by $J_{k+1}, J^-_{k+1} : U \to \mathbb{R}$. We take $u^{-1} := u^0$ for this to be meaningful for $k = 0$. Our main reason for introducing the dependence on u^{k-1} is improve (42) to be explicit in K when it is not affine in y. Along the way we also construct a more conventional inertial method.

5.1 A generalisation of the fundamental theorem

We realign indices to get a simple fundamental condition to verify on each iteration:

Theorem 6 *On a Banach space U, let $H : U \rightrightarrows U^*$, and let $J_k, J^-_k : U \to \overline{\mathbb{R}}$ be Gâteaux-differentiable with the corresponding Bregman divergences $B_k := B_{J_k}$ and $B^-_k := B_{J^-_k}$ for all $k = 1, \ldots, N$. Suppose (IPP) is solvable for $\{u^{k+1}\}_{k\in\mathbb{N}}$ given an initial iterate $u^0 \in U$. If for all $k = 0, \ldots, N-1$, for some $\bar{u} \in U$ and $\mathcal{G}(u^{k+1}, \bar{u}) \in \mathbb{R}$, for all $h^{k+1} \in H(u^{k+1})$ the modified* fundamental condition

$$\langle h^{k+1} | u^{k+1} - \bar{u} \rangle \geq [(B_{k+2} + B^-_{k+3}) - (B_{k+1} + B^-_{k+2})](\bar{u}, u^{k+1}) + \mathcal{G}(u^{k+1}, \bar{u}) \tag{IC}$$

holds, and B^-_{k+1} satisfies the general Cauchy inequality

$$\langle D_1 B^-_{k+1}(u^k, u) | u^k - u' \rangle \leq B'_{k+1}(u^k, u) + B''_{k+1}(u', u^k) \quad (u, u' \in X) \tag{32}$$

for some $B'_{k+1}, B''_{k+1} : U \times U \to \mathbb{R}$, then we have the modified descent inequality

$$[B_{N+1} + B^-_{N+2} - B''_{N+1}](\bar{u}, u^N) + \sum_{k=0}^{N-1} [B_{k+1} + B^-_{k+2} - B''_{k+1} - B'_{k+2}](u^{k+1}, u^k) + \sum_{k=0}^{N-1} \mathcal{G}(u^{k+1}, \bar{u}) \leq [B_1 + B^-_2](\bar{u}, u^0). \tag{ID}$$

Proof We can write (IPP) as

$$0 = h^{k+1} + D_1 B_{k+1}(u^{k+1}, u^k) + D_1 B^-_{k+1}(u^k, u^{k-1}) \text{ for some } h^{k+1} \in H(u^{k+1}). \tag{33}$$

Testing (IPP) by applying $\langle \cdot | u^{k+1} - \bar{u} \rangle$ we obtain

$$0 = \langle h^{k+1} + D_1 B_{k+1}(u^{k+1}, u^k) + D_1 B^-_{k+1}(u^k, u^{k-1}) | u^{k+1} - \bar{u} \rangle.$$

Summing over $k = 0, \ldots, N-1$ and using $u^{-1} = u^0$ to eliminate $B_1^-(u^0, u^{-1}) = 0$, we rearrange

$$0 = S_N + \sum_{k=0}^{N-1} \langle h^{k+1} + D_1[B_{k+1} + B_{k+2}^-](u^{k+1}, u^k) | u^{k+1} - \bar{u} \rangle \tag{34}$$

for

$$S_N := \langle D_1 B_{J_{N+1}^-}(u^N, u^{N-1}) | \bar{u} - u^N \rangle + \sum_{k=0}^{N-1} \langle D_1 B_{J_{k+1}^-}(u^k, u^{k-1}) | u^{k+1} - u^k \rangle.$$

Abbreviating $\bar{B}_{k+1} := B_{k+1} + B_{k+2}^-$ and using (IC) and the three-point identity (6) in (34) we obtain

$$0 \geq S_N + \sum_{k=0}^{N-1} \left(\bar{B}_{k+2}(\bar{u}, u^{k+1}) - \bar{B}_{k+1}(\bar{u}, u^k) + \bar{B}_{k+1}(u^{k+1}, u^k) + \mathcal{G}(u^{k+1}, \bar{u}) \right).$$

Using the generalised Cauchy inequality (32) and, again, that $u^{-1} = u^0$, we get

$$\begin{aligned} S_N &\geq -B'_{N+1}(u^N, u^{N-1}) - B''_{N+1}(\bar{u}, u^N) - \sum_{k=0}^{N-1} \left(B'_{k+1}(u^k, u^{k-1}) + B''_{k+1}(u^{k+1}, u^k) \right) \\ &= -B''_{N+1}(\bar{u}, u^N) - \sum_{k=0}^{N-1} [B''_{k+1} + B'_{k+2}](u^{k+1}, u^k). \end{aligned}$$

These two inequalities yield (ID). □

5.2 Inertia (almost) as usually understood

We take $J_{k+1} = J^0$ and $J_{k+1}^- = -\lambda_k J^0$ for some $\lambda_k \in \mathbb{R}$. We then expand (IPP) as

Inertial PDBS

Iteratively over $k \in \mathbb{N}$, solve for x^{k+1} and y^{k+1}:

$$\begin{aligned} &(1+\lambda_k)[DJ_X(x^k) - D_x K(x^k, y^k)] - \lambda_k[DJ_X(x^{k-1}) - D_x K(x^{k-1}, y^{k-1})] \\ &\qquad \in DJ_X(x^{k+1}) + \partial F(x^{k+1}), \\ &(1+\lambda_k)[DJ_Y(y^k) - D_y K(x^k, y^k)] - \lambda_k[DJ_Y(y^{k-1}) - D_y K(x^{k-1}, y^{k-1})] \\ &\qquad \in DJ_Y(y^{k+1}) + \partial G_*(y^{k+1}) - 2D_y K(x^{k+1}, y^{k+1}) \end{aligned} \tag{35}$$

If X and Y are Hilbert spaces with $J_X = \tau^{-1} N_X$ and $J_Y = \sigma^{-1} N_Y$ the standard generating functions divided by some step length parameters $\tau, \sigma > 0$, and $K(x, y) = \langle Ax|y\rangle$ for $A \in \mathbb{L}(X; Y)$, (35) reduces to the inertial method of [13]:

Inertial PDPS for bilinear K

With initial $\tilde{x}^0 = x^0$ and $\tilde{y}^0 = y^0$, iterate over $k \in \mathbb{N}$:

$$\begin{aligned} x^{k+1} &:= \operatorname{prox}_{\tau F}(\tilde{x}^k - \tau A^* \tilde{y}^k), \\ y^{k+1} &:= \operatorname{prox}_{\sigma G_*}(\tilde{y}^k + \sigma A(2x^{k+1} - \tilde{x}^k)), \\ \tilde{x}^{k+1} &:= (1 + \lambda_{k+1})x^{k+1} - \lambda_{k+1} x^k, \\ \tilde{y}^{k+1} &:= (1 + \lambda_{k+1})y^{k+1} - \lambda_{k+1} y^k. \end{aligned} \tag{36}$$

More generally, however, (35) does not directly apply inertia to the iterates. It applies inertia to K.

The general Cauchy inequality (32) automatically holds by the three-point identity (6) with $J''_{k+1} = J'_{k+1} = J^-_{k+1}$ if $B^-_{k+1} \geq 0$, which is to say that J^-_{k+1} is convex. This is the case if $\lambda_k \leq 0$. For usual inertia we, however, want $\lambda_k > 0$. We will therefore use Lemma 1, requiring:

Assumption 2 For some $\beta > 0$, in a domain $\Omega \subset X \times Y$,

$$|\langle D_1 B^0(u^k, u)|u^k - u\rangle| \leq B^0(u^k, u) + \beta B^0(u', u^k) \quad (u, u', u^k \in \Omega). \tag{37}$$

Moreover, the parameters $\{\lambda_k\}_{k\in\mathbb{N}}$ are non-increasing and for some $\varepsilon > 0$,

$$0 \leq \lambda_{k+1} \leq \frac{1 - \varepsilon - \lambda_k \beta}{2} \quad (k \in \mathbb{N}). \tag{38}$$

Example 22 Suppose the generating function J^0 is γ-strongly subdifferentiable (i.e., B^0 is γ-elliptic, see Sections 4.2 and 4.3) within $\Omega \subset X \times Y$ and satisfies the subdifferential smoothness property (8) with the factor $L > 0$. Then by Lemma 1, (37) holds with $\beta = L\gamma^{-1}$ in some domain $\Omega \subset X \times Y$.

In particular, the subdifferential smoothness property (8) holds in Hilbert spaces if DJ^0 is L-Lipschitz within Ω; see [3, Theorem 18.15] or [39, Appendix C]. Thus, if $J_X = \tau^{-1} N_X$, $J_Y = \sigma^{-1} N_Y$, and DK is L_{DK}-Lipschitz within Ω, we can take $L = \max\{\sigma^{-1}, \tau^{-1}\} + L_{DK}$. We computed L_{DK} for some specific K in Section 4.2.

Example 23 If $K(x, y) = \langle Ax|y\rangle$ with $A \in \mathbb{L}(X; Y^*)$, and if $J_X = \tau^{-1} N_X$, $J_Y = \sigma^{-1} N_Y$,in Hilbert spaces X and Y, then $B^0(u', u) = \frac{1}{2\tau}\|x - x'\|^2 + \frac{1}{2\sigma}\|y - y'\|^2 + \langle A(x - x')|y - y'\rangle$. By standard Cauchy inequality, (37) holds for $\beta = 1$ in $\Omega = X \times Y$. Consequently the next example recovers the upper bound for λ in [13]:

Example 24 The bound (38) holds for some $\varepsilon > 0$ if $\lambda_k \equiv \lambda$ for $0 \leq \lambda < 1/(2 + \beta)$.

Lemma 8 *Suppose Assumption 2 holds and that* (C^2) *holds within* $\Omega_{\bar u}$ *for some* $\bar u \in \Omega$ *and* $\mathcal{G}(u, \bar u)$. *Given* $u^0 \in \Omega$, *suppose the iterates generated by the inertial PDBS* (35) *satisfy* $\{u^k\}_{k=0}^N \subset \Omega_{\bar u} \cap \Omega$. *Then*

$$\varepsilon B^0(\bar u, u^N) + \varepsilon \sum_{k=0}^{N-1} B^0(u^{k+1}, u^k) + \sum_{k=0}^{N-1} \mathcal{G}(u^{k+1}, \bar u) \le (1 - \lambda_1) B^0(\bar u, u^0). \tag{39}$$

Proof Since $B_{k+1} = B^0$ and $B^-_{k+1} = -\lambda_k B^0$ for all $k \in \mathbb{N}$,

$$(B_{k+2} + B^-_{k+3}) - (B_{k+1} + B^-_{k+2}) = (\lambda_{k+1} - \lambda_{k+2}) B^0.$$

Since λ_k is decreasing and B^0 is semi-elliptic within $\Omega \supset \{u^k, \bar u\}$, we deduce that $(\lambda_{k+1} - \lambda_{k+2}) B^0(\bar u, u^k) \ge 0$. Consequently (IC) holds if (C) does. By the proof of Lemma 3, (IC) then holds if (C^2) does. Using (37), (32) holds with $B'_{k+1} = \lambda_k B_0$ and $B''_{k+1} = \lambda_k \beta B_0$. Referring to Theorem 6, we now obtain (ID). We expand

$$[B_{N+1} + B^-_{N+2} - B''_{N+1}](\bar u, u^N) = (1 - \lambda_{k+1} - \lambda_k \beta) B^0(\bar u, u^N) \quad \text{and}$$
$$[B_{k+1} + B^-_{k+2} - B''_{k+1} - B'_{k+2}](u^{k+1}, u^k) = (1 - \lambda_{k+1} - \lambda_k \beta - \lambda_{k+1}) B^0(u^{k+1}, u^k).$$

Since $\bar u, u^k \in \Omega$ for all $k = 0, \ldots, N$, using the ellipticity of B^0 within Ω as well as (38) we now estimate the first from below by $\varepsilon B^0(\bar u, u^N)$ and the second by $\varepsilon B^0(u^{k+1}, u^k)$. Thus (ID) produces (39). □

We may now proceed as in Sections 4.6 and 4.7.

Theorem 7 (Convergence, inertial method) *Theorems 2, 3 and 5 apply to the iterates* $\{u^{k+1}\}_{k \in \mathbb{N}}$ *generated by the inertial PDBS* (35) *if we replace the assumptions of (semi-)ellipticity of* B^0 *with Assumption 2.*

Proof We replace Lemma 3 and (D) by Lemma 8 and (39) in the proofs of Theorems 2, 3 and 5. Observe that Assumption 2 implies that B^0 is (semi-)elliptic. □

Remark 7 The inertial PDPS is improved in [38] to yield *non-ergodic* convergence of the Lagrangian gap. To do the "inertial unrolling" that leads to such estimates, one, however, needs to correct for the anti-symmetry introduced by K into H.

Remark 8 Since Theorem 6 does not provide the quantitative Δ-Féjer motonicity used in Theorem 4, we cannot prove linear convergence using our present simplified "testing" approach lacking the "testing parameters" of [39].

5.3 Improvements to the basic method without dual affinity

We now have the tools to improve the basic PDBS (14) to enjoy prox-simple steps for general K not affine in y. Compared to (12) we amend $J_{k+1} = J^0$ by taking

$$
\begin{aligned}
J_{k+1}(x,y) &:= J_X(x) + J_Y(y) - K(x,y) + 2K(x^{k+1},y) \\
&= J^0(x,y) + 2K(x^{k+1},y).
\end{aligned} \tag{40}
$$

This would be enough for K to be explicit in the algorithm, however, proofs of convergence would practically require G_* to be strongly convex even in the convex-concave case. To fix this, we introduce the inertial term generated by

$$
J^-_{k+1}(u) := [J^0 - J_k](u) = -2K(x^k, y). \tag{41}
$$

As always, we write B_{k+1}, B^0, and B^-_{k+1} for the Bregman divergences generated by J_{k+1}, J^0, and J^-_{k+1}.

Since

$$
D_1[B_{k+1} - B^0](u^k, u^{k-1}) + D_1 B^-_{k+1}(u^k, u^{k-1}) = (0, \tilde{y}^*_{k+1})
$$

for

$$
\tilde{y}^*_{k+1} = 2[D_y K(x^{k+1}, y^{k+1}) - D_y K(x^{k+1}, y^k) - D_y K(x^k, y^k) + D_y K(x^k, y^{k-1})],
$$

the algorithm (IPP) expands similarly to (14) as the

Modified PDBS

Iteratively over $k \in \mathbb{N}$, solve for x^{k+1} and y^{k+1}:

$$
\begin{aligned}
&DJ_X(x^k) - D_x K(x^k, y^k) \in DJ_X(x^{k+1}) + \partial F(x^{k+1}) \quad \text{and} \\
&DJ_Y(y^k) + [2D_y K(x^{k+1}, y^k) + D_y K(x^k, y^k) - 2D_y(x^k, y^{k-1})] \\
&\qquad\qquad \in DJ_Y(y^{k+1}) + \partial G_*(y^{k+1}).
\end{aligned} \tag{42}
$$

The method reduces to the basic PDBS (14) when K is affine in y. In Hilbert spaces X and Y with $J_X = \tau^{-1} N_X$ and $J_Y = \sigma^{-1} N_Y$, we can rearrange (42) as

Modified PDPS

Iterate over $k \in \mathbb{N}$:

$$
\begin{aligned}
x^{k+1} &:= \operatorname{prox}_{\tau F}(x^k - \tau \nabla_x K(x^k, y^k)), \\
y^{k+1} &:= \operatorname{prox}_{\sigma G_*}(y^k + \sigma[2\nabla_y K(x^{k+1}, y^k) + \nabla_y K(x^k, y^k) - 2\nabla_y K(x^k, y^{k-1})]).
\end{aligned} \tag{43}
$$

Remark 9 The modified PDPS (43) is slightly more complicated than the method in [18], which would update

$$
y^{k+1} := \operatorname{prox}_{\sigma G_*}(y^k + \sigma K(2x^{k+1} - x^k, y^k)).
$$

Likewise, (42) is different from the algorithm presented in [26] for convex-concave K. It would, for the standard generating functions, update[6]

$$y^{k+1} := \operatorname{prox}_{\sigma G_*}(y^k + \sigma[2K(x^{k+1}, y^k) - K(x^k, y^{k-1})]).$$

We could produce this method by taking $J^-_{k+1}(u) = -K(x^k, y)$. However, the convergence proofs would require some additional steps.

The main difference to the overall analysis of Section 4 is in bounding from below the Bregman divergences in (ID). We now have

$$B_{N+1} + B^-_{N+2} - B''_{N+1} = B^0 - B''_{N+1} \quad \text{and} \tag{44a}$$

$$B_{k+1} + B^-_{k+2} - B''_{k+1} - B'_{k+2} = B^0 - B''_{k+1} - B'_{k+2}. \tag{44b}$$

If $D_y K(x^k, \cdot)$ is $L_{DK,y}$-Lipschitz,

$$\begin{aligned} \langle D_1 B^-_{k+1}(u^k, u) | u^k - u' \rangle &= 2\langle D_y K(x^k, y^k) - D_y K(x^k, y) | y^k - y' \rangle \\ &\le \sqrt{L_{DK,y}} \|y - y^k\|^2 + \sqrt{L_{DK,y}} \|y' - y^k\|^2 \\ &=: B'_{k+1}(u^k, u) + B''_{k+1}(u', u^k). \end{aligned} \tag{45}$$

Therefore, for the modified descent inequality (ID) to be meaningful, we require:

Assumption 3 We assume that $\|D_y K(x, y) - D_y K(x, y')\| \le L_{DK,y} \|y - y'\|$ when $(x, y), (x, y') \in \Omega$ for some domain $\Omega \subset X \times Y$. Moreover, for some $\varepsilon \ge 0$ we have

$$B^0(u, u') \ge \frac{\varepsilon}{2} \|u - u'\|^2_{X \times Y} + 2\sqrt{L_{DK,y}} \|y - y'\|^2_Y \quad (u, u' \in \Omega). \tag{46}$$

We say that the present assumption holds *strongly* if $\varepsilon > 0$.

Example 25 If K is affine in y, $L_{DK,y} = 0$. Therefore, Assumption 3 reduces to the (semi-)ellipticity of B^0, which can be verified as in Sections 4.2 and 4.3.

Example 26 Generally, it is easy to see that if one of the results of Section 4.2 holds with $\tilde{\sigma} = 1/(\sigma^{-1} - 4\sqrt{L_{DK,y}}) > 0$ in place of σ, then (46) holds. In particular, if K has L_{DK}-Lipschitz derivative within Ω, then Lemma 2 gives the condition $1 \ge L_{DK} \max\{\tau, \sigma/(1 - 4\sigma\sqrt{L_{DK,y}})\}$ and $1 > 4\sigma\sqrt{L_{DK,y}}$ for (46) to hold with $\varepsilon = 0$. The assumption holds strongly if the first inequality is strict.

Similarly to Lemma 8, we now have the following replacement for Lemma 3:

Lemma 9 *Suppose Assumption 3 holds and* (C^2) *holds within* $\Omega_{\bar{u}}$ *for some* $\bar{u} \in X \times Y$ *and* $\mathcal{G}(u, \bar{u})$. *Given* $u^0 \in X \times Y$, *suppose the iterates generated by the modified PDBS* (42) *satisfy* $\{u^k\}_{k=0}^N \subset \Omega_{\bar{u}}$. *Then*

---

[6] Note that [26] uses the historical ordering of the primal and dual updates from [12], prior to the proof-simplifying discovery of the proximal point formulation in [27]. Hence our y^k is their y^{k+1}.

$$\varepsilon B^0(\bar{u}, u^N) + \varepsilon \sum_{k=0}^{N-1} B^0(u^{k+1}, u^k) + \sum_{k=0}^{N-1} \mathcal{G}(u^{k+1}, \bar{u}) \le [B_1 + B_2^-](\bar{u}, u^0). \tag{47}$$

Proof Inserting (40) and (41), (IC) reduces to (C), which follows from (C^2) as in Lemma 3. We verify (32) via (45) and Assumption 3. Thus Theorem 6 proves (ID). Inserting (44) and (46) with B'_{k+1} and B''_{k+1} from (45) into (ID) proves (47). □

We may now proceed as in Sections 4.6 and 4.7 to prove convergence:

Theorem 8 (Convergence, modified method) *Theorems 2, 3 and 5 apply to the iterates $\{u^{k+1}\}_{k\in\mathbb{N}}$ generated by the modified PDBS (42) if we replace the assumptions of semi-ellipticity (resp. ellipticity) of B^0 with Assumption 3 holding (strongly).*

Proof We replace Lemma 3 and (D) by Lemma 9 and (47) in Theorems 2, 3 and 5. Observe that (strong) Assumption 3 implies the (semi-)ellipticity of B^0. □

6 Further directions

We close by briefly reviewing some things not covered, other possible extensions, and alternative algorithms.

6.1 Acceleration

To avoid technical detail, we did not cover $O(1/N^2)$ acceleration. The fundamental ingredients of proof are, however, exactly the same as we have used: sufficient second-order growth and ellipticity of the Bregman divergences B_k^0, which are now iteration-dependent. Additionally, we must use some of that growth to make the B_k^0 grow. For bilinear K in Hilbert spaces, such an argument can be found in [39]; for $K(x, y) = \langle A(x)|y\rangle$ in [17]; and for general K in [18]. As mentioned in Remarks 1 and 9, the algorithms in the latter two differ slightly from the ones presented here.

6.2 Stochastic methods

It is possible to refine the block-adapted (16) and its accelerated version into stochastic methods. The idea is to take on each step subsets of primal-blocks $S(i) \subset \{1, \ldots, m\}$ and dual blocks $V(i+1) \subset \{1, \ldots, n\}$ and to only update the corresponding x_j^{k+1} and y_ℓ^{k+1}. Full discussion of such technical algorithms outside the scope of our present overview. We refer to [40] for an approach covering block-adapted acceleration and both primal- and dual randomisation in the case of bilinear K, but see also [11] for a more basic version. For more general K affine in y, see [32].

6.3 Alternative Bregman divergences

We have used Bregman divergences as a proof tool, in the end opting for the standard quadratic generating functions on Hilbert spaces. Nevertheless, our theory works for arbitrary Bregman divergences. The practical question is whether F and G_* remain prox-simple with respect to such a divergence. This can be the case for the "entropic distance" generated on $L^1(\Omega; [0,\infty))$ by

$$J(x) := \begin{cases} \int_\Omega x(t) \ln x(t) \, dt, & x \geq 0 \text{ a.e. on } \Omega, \\ \infty, & \text{otherwise} \end{cases}$$

See, for example, [10] for a Landweber method (gradient descent on regularised least squares) based on such a distance.

6.4 Alternative approaches

The derivative $D_1 B^0$ in (13) can be seen as a preconditioner, replacing $\tau(u - u')$ in the proximal point method (11). Our choice of B^0 is not the only option.

Consider $\min_{x \in X} E(x) + G(Ax)$, where E is Gâteaux-differentiable and A linear. Classical forward–backward splitting is impractical as $G \circ A$ is in general not prox-simple. Assuming G to have the preconjugate G_*, we can write this problem as an instance of (S) with $F = 0$ and $K(x, y) = E(x) + \langle Ax | y \rangle$. Therefore the methods we have presented are applicable. However, in this instance, also $J^0(u) := \frac{1}{2}\|u\|^2_{X \times Y} + \frac{1}{2}\|A^* y\|^2_{X^*}$ would produce an algorithm with realisable steps. In analogy to the PDPS, it might be called the *primal dual explicit spitting* (PDES). The method was introduced in [30] for $E(z) = \frac{1}{2}\|b - z\|^2$ as the "generalised iterative soft-thresholding" (GIST), but has also been called the *primal-dual fixed point method* (PDFP, [14]) and the *proximal alternating predictor corrector* (PAPC, [23]).

The classical *Augmented Lagrangian* method solves the saddle point problem

$$\min_x \max_y \; F(x) + \frac{\tau}{2}\|E(x)\|^2 + \langle E(x) | y \rangle, \tag{48}$$

alternatingly for x and y. The *alternating directions method of multipliers* (ADMM) of [25, 1] takes $E(x) = Ax_1 + Bx_2 - c$ and $F(x) = F_1(x_1) + F_2(x_2)$ for $x = (x_1, x_2)$, and alternates between solving (48) for x_1, x_2, and y, using the most recent iterate for the other variables. The method cannot be expressed in our Bregman divergence framework, as the preconditioner $D_1 B_{k+1}(\,\cdot\,, x^k)$ would need to be non-symmetric. The steps of the method are potentially expensive, each itself an optimisation problem. Hence the *preconditioned ADMM* of [43], equivalent to the PDPS and the classical *Douglas–Rachford splitting* (DRS, [22]) applied to appropriate problems [12]. The preconditioned ADMM was extended to nonlinear E in [6].

Based on derivations avoiding the Lipschitz gradient assumption (cocoercivity) in forward-backward splitting, [31] moves the over-relaxation step $\bar{x}^{k+1} := 2x^{k+1} - x^k$ of the PDPS outside the proximal operators. This amounts to taking $J_{k+1}^{-} = \lambda_k K$ in Section 5.2 instead of $J_{k+1}^{-}(x, y) = \lambda_k J^0 = \lambda_k[\tau^{-1} J_X(x) + \sigma^{-1} J_Y(y) - K(x, y)]$, so is "partial inertia"; compare the "corrected inertia" of [38].

An *over-relaxed* variant of the same idea maybe found in [7]. We have not discussed over-relaxation of entire algorithms. To briefly relate it to the basic inertia of (36), the latter "rebases" the algorithm at the inertial iterate $\tilde{u}^k$ constructed from u^k and u^{k-1}, whereas over-relaxation would construct $\tilde{u}^k$ from u^k and $\tilde{u}^{k-1}$. The derivation in [7] is based on applying Douglas–Rachford splitting on a lifted problem. The basic over-relaxation of the PDPS is known as the Condat–Vũ method [20, 42].

6.5 Functions on manifolds and Hadamard spaces

The PDPS has been extended in [5] to functions on Riemannian manifolds; the problem $\min_{x \in \mathcal{M}} F(x) + G(Ex)$, where $E : \mathcal{M} \to \mathcal{N}$ with $\mathcal{M}$ and $\mathcal{N}$ Riemannian manifolds. In general, between manifolds, there are no linear maps, so E is nonlinear. Indeed, besides introducing a theory of conjugacy for functions on manifolds, the algorithm presented in [5] is based on the NL-PDPS of [37, 17].

Convergence could only be proved on *Hadamard manifolds*, which are special: a type of three-point inequality holds [21, Lemma 12.3.1]. Indeed, in even more general *Hadamard spaces* with the metric d, for any three points $x^{k+1}, x^k, \bar{x}$, we have [2, Corollary 1.2.5]

$$\frac{1}{2} d(x^k, x^{k+1})^2 + \frac{1}{2} d(x^{k+1}, \bar{x})^2 - \frac{1}{2} d(x^k, \bar{x})^2 \le d(x^k, x^{k+1}) d(\bar{x}, x^{k+1}). \tag{49}$$

Therefore, given a function f on such a space, to derive a simple proximal point algorithm, having constructed the iterate x^k we might try to find x^{k+1} such that

$$f(x^{k+1}) + d(x^k, x^{k+1}) \le f(x^k).$$

Multiplying this inequality by $d(\bar{x}, x^{k+1})$ and using the three-point inequality (49),

$$\frac{1}{2} d(x^k, x^{k+1})^2 + \frac{1}{2} d(x^{k+1}, \bar{x})^2 + [f(x^{k+1}) - f(x^k)] d(\bar{x}, x^{k+1}) \le \frac{1}{2} d(x^k, \bar{x})^2.$$

If the space is bounded, $d(\bar{x}, x^{k+1}) \le C$, so since $f(x^k) \ge f(x^{k+1})$, we may telescope and proceed as before to obtain convergence.

The Hadamard assumption is restrictive: if a Banach space is Hadamard, it is Hilbert, while a Riemannian manifold is Hadamard if it is simply connected with a non-positive sectional curvature [2, section 1.2].

Acknowledgements Academy of Finland grants 314701 and 320022.

References

1. K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*, Stanford University Press, 1958.
2. M. Bačák, *Convex Analysis and Optimization in Hadamard Spaces*, Nonlinear Analysis and Applications, De Gruyter, 2014.
3. H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, 2 edition, 2017, doi:10.1007/978-3-319-48311-5.
4. A. Beck, *First-Order Methods in Optimization*, SIAM, 2017, doi:10.1137/1.9781611974997.
5. R. Begmann, R. Herzog, D. Tenbrick, and J. Vidal-Núñez, Fenchel duality for convex optimization and a primal dual algorithm on Riemannian manifolds, 2019, arXiv:1908.02022.
6. M. Benning, F. Knoll, C. B. Schönlieb, and T. 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*, 2016, 117–126, doi:10.1007/978-3-319-55795-3_10, arXiv:1511.00425.
7. K. Bredies and H. Sun, Preconditioned Douglas–Rachford splitting methods for convex-concave saddle-point problems, *SIAM Journal on Numerical Analysis* 53 (2015), 421–444, doi:10.1137/140965028.
8. H. Brezis, M. G. Crandall, and A. Pazy, Perturbations of nonlinear maximal monotone sets in Banach space, *Communications on Pure and Applied Mathematics* 23 (1970), 123–144, doi:10.1002/cpa.3160230107.
9. F. E. Browder, Convergence theorems for sequences of nonlinear operators in Banach spaces, *Mathematische Zeitschrift* 100 (1967), 201–225, doi:10.1007/bf01109805.
10. M. Burger, E. Resmerita, and M. Benning, An entropic Landweber method for linear ill-posed problems, 2019, arXiv:1906.10032.
11. A. Chambolle, M. Ehrhardt, P. Richtárik, and C. Schönlieb, Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications, *SIAM Journal on Optimization* 28 (2018), 2783–2808, doi:10.1137/17m1134834.
12. A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision* 40 (2011), 120–145, doi:10.1007/s10851-010-0251-1.
13. A. Chambolle and T. Pock, On the ergodic convergence rates of a first-order primal–dual algorithm, *Mathematical Programming* (2015), 1–35, doi:10.1007/s10107-015-0957-3.
14. P. Chen, J. Huang, and X. Zhang, A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration, *Inverse Problems* 29 (2013), 025011, doi:10.1088/0266-5611/29/2/025011.
15. G. Chierchia, E. Chouzenoux, P. L. Combettes, and J. C. Pesquet, The Proximity Operator Repository, 2019, http://proximity-operator.net. Online resource.
16. F. Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, 1990, doi:10.1137/1.9781611971309.
17. C. Clason, S. Mazurenko, and T. 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.
18. C. Clason, S. Mazurenko, and T. Valkonen, Primal–dual proximal splitting and generalized conjugation in non-smooth non-convex optimization, 2019, arXiv:1901.02746. Submitted.
19. C. Clason and T. 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.
20. L. Condat, A Primal–Dual Splitting Method for Convex Optimization Involving Lipschitzian, Proximable and Linear Composite Terms, *Journal of Optimization Theory and Applications* 158 (2013), 460–479, doi:10.1007/s10957-012-0245-9.
21. M. P. do Carmo, *Riemannian Geometry*, Mathematics: Theory & Applications, Birkhäuser, 2013.

22. J. Douglas, Jim and J. Rachford, H. H., On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables, *Transactions of the American Mathematical Society* 82 (1956), 421–439, doi:10.2307/1993056.
23. Y. Drori, S. Sabach, and M. 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.
24. I. Ekeland and R. Temam, *Convex analysis and variational problems*, SIAM, 1999.
25. D. Gabay, Applications of the Method of Multipliers to Variational Inequalities, in *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, M. Fortin and R. Glowinski (eds.), volume 15 of Studies in Mathematics and its Applications, North-Holland, 1983, 299–331.
26. E. Y. Hamedani and N. S. Aybat, A primal-dual algorithm for general convex-concave saddle point problems, 2018, arXiv:1803.01401.
27. B. He and X. Yuan, Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective, *SIAM Journal on Imaging Sciences* 5 (2012), 119–149, doi:10.1137/100814494.
28. J. B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Grundlehren Text Editions, Springer, 2004.
29. T. Hohage and C. Homann, A Generalization of the Chambolle-Pock Algorithm to Banach Spaces with Applications to Inverse Problems, 2014, arXiv:1412.0126.
30. I. Loris and C. Verhoeven, On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty, *Inverse Problems* 27 (2011), 125007, doi:10.1088/0266-5611/27/12/125007.
31. Y. Malitsky and M. K. Tam, A forward-backward splitting method for monotone inclusions without cocoercivity, 2018, arXiv:1808.04162.
32. S. Mazurenko, J. Jauhiainen, and T. Valkonen, Primal–dual block-proximal splitting for a class of non-convex problems, 2019. Work being finalised.
33. G. J. Minty, On the Maximal domain of a “monotone” function, *The Michigan Mathematical Journal* 8 (1961), 135—137.
34. Z. Opial, Weak convergence of the sequence of successive approximations for nonexpansive mappings, *Bulletin of the American Mathematical Society* 73 (1967), 591–597, doi:10.1090/s0002-9904-1967-11761-0.
35. T. Pock, D. Cremers, H. Bischof, and A. Chambolle, An algorithm for minimizing the Mumford-Shah functional, in *12th IEEE Conference on Computer Vision*, IEEE, 2009, 1133–1140, doi:10.1109/iccv.2009.5459348.
36. R. T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM Journal on Optimization* 14 (1976), 877–898, doi:10.1137/0314056.
37. T. 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.
38. T. Valkonen, Inertial, corrected, primal–dual proximal splitting, 2018, arXiv:1804.08736. Submitted.
39. T. 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.
40. T. Valkonen, Block-proximal methods with spatially adapted acceleration, *Electronic Transactions on Numerical Analysis* 51 (2019), 15–49, doi:10.1553/etna_vol51s15, arXiv:1609.07373.
41. T. Valkonen and T. 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.
42. B. C. Vũ, A splitting algorithm for dual monotone inclusions involving cocoercive operators, *Advances in Computational Mathematics* 38 (2013), 667–681, doi:10.1007/s10444-011-9254-8.
43. X. Zhang, M. Burger, and S. Osher, A unified primal-dual algorithm framework based on Bregman iteration, *Journal of Scientific Computing* 46 (2011), 20–46, doi:10.1007/s10915-010-9408-8.