Electronic Transactions on Numerical Analysis.
Volume 51, pp. 15–49, 2019.

ISSN 1068–9613.
DOI: 10.1553/etna_vol51s15

ETNA
Kent State University and
Johann Radon Institute (RICAM)

BLOCK-PROXIMAL METHODS WITH SPATIALLY ADAPTED ACCELERATION*

TUOMO VALKONEN†

Abstract. We study and develop (stochastic) primal-dual block-coordinate descent methods for convex problems based on the method due to Chambolle and Pock. Our methods have known convergence rates for the iterates and the ergodic gap of $O(1/N^2)$ if each block is strongly convex, $O(1/N)$ if no convexity is present, and more generally a mixed rate $O(1/N^2) + O(1/N)$ for strongly convex blocks if only some blocks are strongly convex. Additional novelties of our methods include blockwise-adapted step lengths and acceleration as well as the ability to update both the primal and dual variables randomly in blocks under a very light compatibility condition. In other words, these variants of our methods are doubly-stochastic. We test the proposed methods on various image processing problems, where we employ pixelwise-adapted acceleration.

Key words. PDHGM, Chambolle–Pock method, stochastic, doubly-stochastic, blockwise, primal-dual

AMS subject classifications. 49M29, 65K10, 65K15, 90C30, 90C47

1. Introduction.

We want to efficiently solve optimisation problems of the form

$$\min_x \; G(x) + F(Kx), \tag{P$_0$}$$

arising, in particular, from image processing and inverse problems. We assume $G : X \to \overline{\mathbb{R}}$ and $F : Y \to \overline{\mathbb{R}}$ to be convex, proper, and lower semicontinuous on Hilbert spaces X and Y and $K \in \mathcal{L}(X;Y)$ to be a bounded linear operator. We are particularly interested in block-separable functionals

$$G(x) = \sum_{j=1}^{m} G_j(P_j x) \qquad \text{and} \qquad F^*(y) = \sum_{\ell=1}^{n} F_\ell^*(Q_\ell y), \tag{GF}$$

where F^* is the Fenchel conjugate of F. The operators $P_1, \dots, P_m$ are projections in X with $\sum_{j=1}^m P_j = I$ and $P_j P_i = 0$ if $i \neq j$. Likewise, $Q_1, \dots, Q_n$ are projection operators in Y. We assume all the component functions G_j and F_ℓ^* to be convex, proper, and lower semicontinuous, and the subdifferential sum rule to hold for the expressions (GF).

Several first-order optimisation methods have been developed for (P$_0$) without block-separable structure, typically for both G and F convex and K linear. Recently also some non-convexity and non-linearity has been introduced [4, 21, 23, 37]. In applications in image processing and data science, one of G or F is typically non-smooth. Effective algorithms operating directly on the primal problem (P$_0$), or its dual, therefore tend to be a form of classical forward-backward splitting, occasionally called iterative soft-thresholding [1, 12].

In big data optimisation, several forward-backward block-coordinate descent methods have been developed for (P$_0$) with block-separable G. In each step, the methods update only a random subset of blocks $x_j := P_j x$ in parallel; see the review [39] and the original articles [2, 9, 11, 16, 22, 25, 28, 29, 30, 31, 42]. Typically F is assumed smooth, and often, each G_j is strongly convex. Besides parallelism, an advantage of these methods is the exploitation of *blockwise* factors of smoothness and strong convexity. These can help convergence by being better than the global factor.

*Received February 26, 2018. Accepted January 3, 2019. Published online on March 1, 2019. Recommended by Fiorella Sgallari.

†ModeMat, Escuela Politécnica Nacional, Quito, Ecuador; *previously* Department of Mathematical Sciences, University of Liverpool, United Kingdom (tuomo.valkonen@iki.fi).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Unfortunately, primal-only and dual-only stochastic methods, as discussed above, are rarely applicable to image processing problems. These and many other problems do not satisfy the assumed separability and smoothness assumptions. On the other hand, an additional Moreau-Yosida (aka. Huber, aka. Nesterov) regularisation of the problem, which would provide the required smoothness, would alter the problem, losing the essential non-smooth characteristics. Generally, even without the splitting of the problem into blocks and the introduction of stochasticity, primal-only or dual-only methods can be inefficient on more complicated problems. Proximal steps, which are typically used to deal with non-smooth components of the problem, can in particular be as expensive as the original optimisation problems itself. In order to make these steps cheap, the problem has to be formulated appropriately. Such a reformulation can often be provided through primal-dual approaches.

With the Fenchel conjugate F^*, we can write (P_0) as

$$\min_x \max_y \; G(x) + \langle Kx, y\rangle - F^*(y).$$

The method of Chambolle and Pock [6, 27] is popular for this formulation. It is also called the PDHGM (Primal-Dual Hybrid Gradient Method, Modified) in [14] and the PDPS (Primal–Dual Proximal Splitting) in [34]. It consists of alternating proximal steps in x and y combined with an over-relaxation step to ensure convergence. The method is closely related to the classical ADMM and Douglas-Rachford splitting. The acronym PDHGM arises from the earlier PDHG [43] that is convergent only in special cases [18]. These connections are discussed in [14].

While early block-coordinate methods only worked with a primal or a dual variable, recently stochastic primal-dual approaches based on the ADMM and the PDHGM have been proposed [3, 15, 24, 26, 33, 40, 41]. Moreover, variants of the ADMM that deterministically update multiple blocks in parallel and afterwards combine the results for the Lagrange multiplier update have been introduced [20]. As with the primal- or dual-only methods, these algorithms can improve convergence by exploiting local properties of the problem. Besides [33, 40, 41], which have restrictive smoothness and strong convexity requirements, little is known about convergence rates.

In this paper, we will derive block-coordinate descent variants of the PDHGM with known convergence rates: $O(1/N^2)$ if each G_j is strongly convex, $O(1/N)$ without any strong convexity, and mixed $O(1/N^2) + O(1/N)$ if some of the G_j are strongly convex. These rates apply to an ergodic duality gap and strongly convex blocks of the iterates. Our methods have the novelty of blockwise-adapted step lengths. In the imaging applications of Section 5 we will even employ pixelwise-adapted step lengths. Moreover, we can update random subsets of *both* primal and dual blocks under a light "nesting condition" on the sampling scheme. Such "doubly-stochastic" updates have previously been possible only in very limited settings [40].

Our present paper is based on [37] on the acceleration of the PDHGM when G is strongly convex only on a subspace: the deterministic two-block case $m = 2$ and $n = 1$ of (GF). Besides enabling (doubly-)stochastic updates and an arbitrary number of both primal *and* dual blocks, in the present work, we derive simplified step length rules through a more careful analysis.

The more abstract basis of our present work has been described in [35]. There we study preconditioning of abstract proximal point methods and "testing" by suitable operators as means of obtaining convergence rates. We recall the relevant aspects of this theory through the course of Sections 2 and 3. In the first of these sections, we start by going through the notation and previous research on the PDHGM in more detail. Then we develop the rough structure of our proposed method. This will depend on several structural conditions that we introduce in Section 2. Afterwards in Section 3 we develop convergence estimates based on technical

ETNA
Kent State University and
Johann Radon Institute (RICAM)

conditions on the various step length and testing parameters. These conditions need to be verified through the development of explicit parameter update rules. We do this in Section 4 along with proving the claimed convergence rates (Theorem 4.5 and its corollaries). We also present there the final detailed versions of our proposed algorithms: Algorithm 1 (doubly stochastic) and Algorithm 2 (simplified). We finish with numerical experiments in Section 5.

2. Background and overall structure of the algorithm. To make the notation definite, we write $\mathcal{L}(X;Y)$ for the space of bounded linear operators between Hilbert spaces X and Y. The identity operator we denote by I. For $T, S \in \mathcal{L}(X;X)$, we use $T \geq S$ to indicate that $T - S$ is positive semi-definite; in particular $T \geq 0$ means that T is positive semi-definite. Also for possibly non-self-adjoint T, we introduce the inner product and norm-like notations

$$\langle x, z\rangle_T := \langle Tx, z\rangle \qquad \text{and} \qquad \|x\|_T := \sqrt{\langle x, x\rangle_T},$$

the latter only defined for positive semi-definite T. We write $T \simeq T'$ if $\langle x, x\rangle_{T'-T} = 0$ for all x.

We denote by $\mathcal{C}(X)$ the set of convex, proper, lower semicontinuous functionals from a Hilbert space X to $\overline{\mathbb{R}} := [-\infty, \infty]$. With $G \in \mathcal{C}(X)$, $F^* \in \mathcal{C}(Y)$, and $K \in \mathcal{L}(X;Y)$, we then wish to solve the minimax problem

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

assuming the existence of a solution $\widehat{u} = (\widehat{x}, \widehat{y})$ that satisfies the optimality conditions

$$-K^*\widehat{y} \in \partial G(\widehat{x}) \qquad \text{and} \qquad K\widehat{x} \in \partial F^*(\widehat{y}). \tag{OC}$$

For the stochastic aspects of our work, we denote by $(\Omega, \mathcal{O}, \mathbb{P})$ the *probability space* consisting of the set Ω of possible realisation of a random experiment, by $\mathcal{O}$ a σ-algebra on Ω, and by $\mathbb{P}$ a probability measure on $(\Omega, \mathcal{O})$. We denote the expectation corresponding to $\mathbb{P}$ by $\mathbb{E}$, the conditional probability with respect to a sub-σ-algebra $\mathcal{O}' \subset \mathcal{O}$ by $\mathbb{P}[\,\cdot\,|\mathcal{O}']$, and the conditional expectation by $\mathbb{E}[\,\cdot\,|\mathcal{O}']$. We refer to [32] for more details.

We also use the following non-standard notation: If $\mathcal{O}$ is a σ-algebra on the space Ω, we denote by $\mathcal{R}(\mathcal{O};V)$ the space of V-valued random variables A such that $A : \Omega \to V$ is $\mathcal{O}$-measurable.

2.1. Preconditioned proximal point methods. Testing for rates. We use the notation

$$u = (x, y)$$

to combine the primal variable x and the dual variable y into a single variable u. Following [19, 37], the primal-dual method of Chambolle and Pock [6] (PDHGM) may then be written in proximal point form as

$$0 \in H(u^{i+1}) + L_i(u^{i+1} - u^i) \tag{PP$_0$}$$

for a monotone operator H encoding the optimality conditions (OC) as $0 \in H(\widehat{u})$ and a *preconditioning* or *step length operator* $L_i = L_i^0$. These are

$$H(u) := \begin{bmatrix} \partial G(x) + K^*y \\ \partial F^*(y) - Kx \end{bmatrix} \qquad \text{and} \qquad L_i^0 := \begin{bmatrix} \tau_i^{-1} & -K^* \\ -\omega_i K & \sigma_{i+1}^{-1} \end{bmatrix}.$$

Here $\tau_i, \sigma_{i+1} > 0$ are step length parameters, and $\omega_i > 0$ are over-relaxation parameters. In the basic version of the algorithm we set $\omega_i = 1$, $\tau_i \equiv \tau_0$, and $\sigma_i \equiv \sigma_0$, assuming

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$\tau_0\sigma_0\|K\|^2 < 1$. Observe that we may equivalently parametrise the algorithm by τ_0 and $\delta = 1 - \|K\|^2\tau_0\sigma_0 > 0$. The method has $O(1/N)$ rate for the ergodic duality gap, which we will return to in Section 3.1.

If G is strongly convex with factor $\gamma > 0$, we may, for $\widetilde{\gamma} \in (0, \gamma]$, accelerate

$$\omega_i := 1/\sqrt{1 + 2\widetilde{\gamma}\tau_i}, \qquad \tau_{i+1} := \tau_i\omega_i, \qquad \text{and} \qquad \sigma_{i+1} := \sigma_i/\omega_i. \tag{2.1}$$

This gives $O(1/N^2)$ convergence of $\|x^N - \widehat{x}\|^2$ to zero. If $\widetilde{\gamma} \in (0, \gamma/2]$, we also obtain $O(1/N^2)$ convergence of an ergodic duality gap.

In [37], we extended the PDHGM to partially strongly convex problems: in (GF) this corresponded to the primal two-block and dual single-block case $m = 2$ and $n = 1$ with only G_1 assumed strongly convex. This extension was based on taking in (PP_0) the preconditioner

$$L_i = \begin{bmatrix} T_i^{-1} & -K^* \\ -\omega_i K & \Sigma_{i+1}^{-1} \end{bmatrix} \tag{2.2}$$

for *invertible* $T_i = \tau_{1,i}P_1 + \tau_{2,i}P_2 \in \mathcal{L}(X;X)$ and $\Sigma_{i+1} = \sigma_{i+1}I \in \mathcal{L}(Y;Y)$. After simple rearrangements of (PP_0), the resulting algorithm could be written more explicitly as

$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - T_iK^*y^i), \tag{2.3a}$$

$$y^{i+1} := (I + \Sigma_{i+1}\partial F^*)^{-1}(y^i + \Sigma_{i+1}K((1+\omega_i)x^{i+1} - \omega_i x^i)). \tag{2.3b}$$

Since G is assumed separable, the first, primal update splits into separate updates for $x_1^{i+1} := P_1x^{i+1}$ and $x_2^{i+1} := P_2x^{i+1}$. Note that this explicit form of the algorithm does not require T_i and Σ_{i+1} to be invertible unlike (PP_0) with the choice (2.2), so this suggests that we could develop stochastic methods that randomly choose one, two, or no primal blocks to update.

To study convergence, it is, however, more practical to work with implicit formulations such as (PP_0). We will shortly see how this works. To make (PP_0) work with non-invertible T_i and Σ_{i+1}, let us reformulate it slightly. In fact, let us define

$$\begin{aligned} W_{i+1} &:= \begin{bmatrix} T_i & 0 \\ 0 & \Sigma_{i+1} \end{bmatrix} \qquad \text{and (for the moment)} \\ M_{i+1} &= \begin{bmatrix} I & -T_iK^* \\ -\widetilde{\omega}_i\Sigma_{i+1}K & I \end{bmatrix}. \end{aligned} \tag{2.4}$$

With this, whether or not T_i and Σ_{i+1} are invertible, (2.3) can be written as the preconditioned proximal point iteration

$$W_{i+1}H(u^{i+1}) + M_{i+1}(u^{i+1} - u^i) \ni 0. \tag{PP}$$

This will be the abstract form of the algorithms that we will develop, however, with the exact form of T_{i+1}, Σ_{i+1}, and M_{i+1} still to be refined.

To study the convergence of (PP), we apply to the *testing* framework introduced in [35, 37]. The idea is to apply $\langle \cdot, u^{i+1} - \widehat{u}\rangle_{Z_{i+1}}$ with a *testing operator* Z_{i+1} to (PP) to "test" it. Thus

$$0 \in \langle W_{i+1}H(u^{i+1}) + M_{i+1}(u^{i+1} - u^i), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}}. \tag{2.5}$$

We need $Z_{i+1}M_{i+1}$ to be self-adjoint and positive semi-definite. This guarantees that $Z_{i+1}M_{i+1}$ can be used to form the local semi-norm $\|\cdot\|_{Z_{i+1}M_{i+1}}$. Indeed, assuming for some linear operator Ξ_{i+1} that H has the operator-relative (strong) monotonicity property

$$\langle H(u') - H(u), u' - u\rangle_{Z_{i+1}W_{i+1}} \geq \|u - u'\|^2_{Z_{i+1}\Xi_{i+1}} \qquad (u, u' \in X \times Y), \tag{2.6}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

then a simple application of the Pythagoras identity

$$\langle u^{i+1}-u^i, u^{i+1}-\widehat{u}\rangle_{Z_{i+1}M_{i+1}} = \frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} - \frac{1}{2}\|u^i-\widehat{u}\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{Z_{i+1}M_{i+1}}$$

yields

$$\frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{Z_{i+1}(M_{i+1}+2\Xi_{i+1})} + \frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} \leq \frac{1}{2}\|u^i-\widehat{u}\|^2_{Z_{i+1}M_{i+1}}.$$

If $Z_{i+2}M_{i+2} \leq Z_{i+1}(M_{i+1}+2\Xi_{i+1})$ for all i, then summing over $i = 0, \dots, N-1$ gives

$$\frac{1}{2}\|u^N-\widehat{u}\|^2_{Z_{N+1}M_{N+1}} + \sum_{i=0}^{N-1}\frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} \leq \frac{1}{2}\|u^0-\widehat{u}\|^2_{Z_1M_1}. \tag{2.7}$$

We therefore see that $Z_{i+1}M_{i+1}$ measures the rates of convergence of the iterates. If our iterations are stochastic, then to obtain deterministic estimates, we can simply take the expectation in (2.7). However, to obtain estimates on a duality gap, we need to work significantly more. We will, therefore, in the beginning of Section 3, after introducing all the relevant concepts and finalising the setup for the present work, quote the appropriate results from [35].

2.2. Stochastic and deterministic block updates. We want to update any subset of any number of primal and dual blocks stochastically. Compatible with the separable structure (GF) of G and F^*, we therefore construct—from individual (possibly random) step length and testing parameters $\tau_{j,i}, \sigma_{\ell,i+1} \geq 0$ and $\phi_{j,i}, \psi_{\ell,i+1} > 0$ as well as random subsets $S(i) \subset \{1,\dots,m\}$ and $V(i+1) \subset \{1,\dots,n\}$—the step length and testing operators

$$W_{i+1} := \begin{bmatrix} T_i & 0 \\ 0 & \Sigma_{i+1} \end{bmatrix} \quad \text{and} \quad Z_{i+1} := \begin{bmatrix} \Phi_i & 0 \\ 0 & \Psi_{i+1} \end{bmatrix} \quad \text{for} \tag{S.a}$$

$$T_i := \sum_{j\in S(i)} \tau_{j,i}P_j, \qquad \Sigma_{i+1} := \sum_{\ell\in V(i+1)} \sigma_{\ell,i+1}Q_\ell, \tag{S.b}$$

$$\Phi_i := \sum_{j=1}^{m} \phi_{j,i}P_j, \qquad \Psi_{i+1} := \sum_{\ell=1}^{n} \psi_{\ell,i+1}Q_\ell \quad (i \geq 0). \tag{S.c}$$

We moreover take as the preconditioner

$$M_{i+1} := \begin{bmatrix} I & -\Phi_i^{-1}\Lambda_i^* \\ -\Psi_{i+1}^{-1}\Lambda_i & I \end{bmatrix} \quad \text{for} \quad \Lambda_i := K\mathring{T}_i^*\Phi_i^* - \Psi_{i+1}\mathring{\Sigma}_{i+1}K \quad \text{with} \tag{S.d}$$

$$\mathring{T}_i := \sum_{j\in\mathring{S}(i)} \tau_{j,i}P_j, \qquad \mathring{S}(i) \subset S(i), \tag{S.e}$$

$$\mathring{\Sigma}_{i+1} := \sum_{\ell\in\mathring{V}(i+1)} \sigma_{\ell,i+1}Q_\ell, \qquad \mathring{V}(i+1) \subset V(i+1). \tag{S.f}$$

The subsets $S(i)$ and $V(i+1)$ are the indices of the blocks

$$x_j := P_j x \quad \text{and} \quad y_\ell := Q_\ell y \tag{2.8}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

of the variables x and y that are to be updated at iteration i.[1] Hence T_i and Σ_{i+1} will not be invertible unless we update all the blocks. Clearly Φ_i, Ψ_{i+1}, T_i, and Σ_{i+1} are self-adjoint and positive semi-definite with Φ_i and Ψ_{i+1} invertible. The subsets $\mathring{S}(i)$ and $\mathring{V}(i+1)$ indicate those blocks of x^{i+1} and of y^{i+1} that are to be updated "independently" of the other variable. We will explain these subsets and the choice of Λ_i in more detail in Section 2.3.

The iterate $u^{i+1} = (x^{i+1}, y^{i+1})$ has to be computable based on a random sampling at iteration i and the information gathered (random variable realisations) before commencing the iteration. For the algorithm to be realisable, it cannot depend on the future. We therefore need to be explicit about the space of each random variable. We model the information available just before commencing iteration i by the σ-algebra $\mathcal{O}_{i-1}$. Thus $\mathcal{O}_{i-1} \subset \mathcal{O}_i$. More precisely, $\mathcal{O}_i$ is the smallest sub-σ-algebra of $\mathcal{O}$ satisfying for all $k = 0, \dots, i$, $j = 1, \dots, m$, and $\ell = 1, \dots, n$ that

$$
\begin{aligned}
&(\mathcal{R}.\text{a}) && \tau_{j,k}, \sigma_{\ell,k+1} \in \mathcal{R}(\mathcal{O}_i; [0,\infty)), && \phi_{j,k}, \psi_{\ell,k+1} \in \mathcal{R}(\mathcal{O}_i; (0,\infty)),\\
&(\mathcal{R}.\text{b}) && S(k) \in \mathcal{R}(\mathcal{O}_i; \mathcal{P}(\{1,\dots,m\})), && V(k+1) \in \mathcal{R}(\mathcal{O}_i; \mathcal{P}(\{1,\dots,n\})),\\
&(\mathcal{R}.\text{c}) && \mathring{S}(k) \in \mathcal{R}(\mathcal{O}_i; \mathcal{P}(\{1,\dots,m\})), && \mathring{V}(k+1) \in \mathcal{R}(\mathcal{O}_i; \mathcal{P}(\{1,\dots,n\})).
\end{aligned}
$$

Here and only here $\mathcal{P}$ denotes the power set. Any other variables can only be random by being constructed from these variables. We thus deduce from (S) and (PP) that

$$
\begin{aligned}
&T_k \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(X;X)), && \Phi_k \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(X;X)), && x^{i+1} \in \mathcal{R}(\mathcal{O}_i; X),\\
&\Sigma_{k+1} \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(Y;Y)), && \Psi_{k+1} \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(Y;Y)), && y^{i+1} \in \mathcal{R}(\mathcal{O}_i; Y).
\end{aligned}
$$

We will also need to assume the *nesting conditions* on sampling,

$$
\begin{aligned}
&(\mathcal{V}.\text{a}) && \mathcal{V}(\mathring{S}(i)) \cap \mathring{V}(i+1) = \emptyset, && \mathcal{V}(S(i) \setminus \mathring{S}(i)) \cap (V(i+1) \setminus \mathring{V}(i+1)) = \emptyset,\\
&(\mathcal{V}.\text{b}) && \mathring{S}(i) \cup \mathcal{V}^{-1}(\mathring{V}(i+1)) \subset S(i), && \mathring{V}(i+1) \cup \mathcal{V}(\mathring{S}(i)) \subset V(i+1),
\end{aligned}
$$

where the set

$$
\mathcal{V}(j) := \{\ell \in \{1,\dots,n\} \mid Q_\ell K P_j \neq 0\}
$$

consists of the dual blocks that are "connected" by K to the primal block with index j. Vice versa, $\mathcal{V}^{-1}(\ell)$ consists of the primal blocks that are "connected" by K to the dual block with index ℓ. Thus ($\mathcal{V}$.b) states that the independent updates (i.e., $\mathring{S}(i)$ and $\mathring{V}(i+1)$) must propagate from primal to dual and vice versa as non-independent updates (i.e., $S(i)$ and $V(i+1)$). The condition ($\mathcal{V}$.a) restricts connections between primal and dual updates: the first part means that the independently updated blocks cannot be connected and by the second part, neither can the non-independent updates. If we use ($\mathcal{V}$.b) as an equality to define $S(i)$ and $V(i+1)$, then the second part of ($\mathcal{V}$.a) holds if $\mathcal{V}(\mathcal{V}^{-1}(\mathring{V}(i+1))) \cap \mathcal{V}(\mathring{S}(i)) = \emptyset$, that is, the condition restricts second-degree connections between the independently updated blocks.

To facilitate referring to all the above structural conditions, we introduce:

ASSUMPTION 2.1 (main structural condition). We assume the structure (GF) and (S) with the limitations ($\mathcal{R}$) and ($\mathcal{V}$) on randomness.

Clearly,

$$
Z_{i+1} M_{i+1} = \begin{bmatrix} \Phi_i & -\Lambda_i^* \\ -\Lambda_i & \Psi_{i+1} \end{bmatrix} \tag{2.9}
$$

[1] The iteration index is off-by-one for $\sigma_{\ell,i+1}$ and $\psi_{\ell,i+1}$ for reasons of the historical development of the Chambolle-Pock method, when it was not written as a preconditioned proximal point method.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

is self-adjoint. We need to prove that it is positive semi-definite. We will do this in Section 4 using the functions κ_ℓ introduced next. We show afterwards in Example 2.3 that these functions are a block structure-adapted generalisation of the simple bound $K \leq \|K\| I$.

DEFINITION 2.2 *Let $\mathcal{P} := \{P_1, \dots, P_m\}$ and $\mathcal{Q} := \{Q_1, \dots, Q_n\}$. We write $(\kappa_1, \dots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$ if each $\kappa_\ell : [0,\infty)^m \to [0,\infty)$ is monotone ($\ell = 1, \dots, n$) and the following hold:*

(i) (Estimation) For all $(z_{\ell,1}, \dots, z_{\ell,m}) \subset [0,\infty)^m$ and $\ell = 1, \dots, n$,

$$\sum_{j=1}^{m} \sum_{\ell,k=1}^{n} z_{\ell,j}^{1/2} z_{k,j}^{1/2} Q_\ell K P_j K^* Q_k \leq \sum_{\ell=1}^{n} \kappa_\ell(z_{\ell,1}, \dots, z_{\ell,m}) Q_\ell.$$

(ii) (Boundedness) For some $\overline{\kappa} > 0$ and all $(z_1, \dots, z_m) \subset [0,\infty)^m$,

$$\kappa_\ell(z_1, \dots, z_m) \leq \overline{\kappa} \sum_{j=1}^{m} z_j.$$

(iii) (Non-degeneracy) There exists $\underline{\kappa} > 0$, and for all $j = 1, \dots, m$ a choice of $\ell^*(j) \in \{1, \dots, n\}$ exists such that for all $(z_1, \dots, z_m) \subset [0,\infty)^m$,

$$\underline{\kappa} z_j \leq \kappa_{\ell^*(j)}(z_1, \dots, z_m) \qquad (j = 1, \dots, m).$$

The choice of κ allows us to construct different algorithms. Here we consider a few possibilities, first a simple one, and then a more optimal one.

EXAMPLE 2.3 (Worst-case κ). We may estimate

$$\sum_{j=1}^{m} \sum_{\ell,k=1}^{n} z_{\ell,j}^{1/2} z_{k,j}^{1/2} Q_\ell K P_j K^* Q_k \leq \sum_{\ell,k=1}^{n} \overline{z}_\ell^{1/2} \overline{z}_k^{1/2} Q_\ell K K^* Q_k \leq \sum_{\ell=1}^{n} \overline{z}_\ell \|K\|^2 Q_\ell.$$

Therefore Definition 2.2(i) and (ii) hold with $\overline{\kappa} = \|K\|^2$ for the monotone choice

$$\kappa_\ell(z_1, \dots, z_m) := \|K\|^2 \max\{z_1, \dots, z_m\}.$$

Clearly also $\underline{\kappa} = \overline{\kappa}$ for any choice of $\ell^*(j) \in \{1, \dots, n\}$. This choice of κ_ℓ corresponds to the rule $\tau\sigma\|K\|^2 < 1$ in the PDHGM method.

EXAMPLE 2.4 (Balanced κ). Take minimal κ_ℓ that satisfy Definition 2.2 and the balancing condition $\kappa_\ell(z_{\ell,1}, \dots, z_{\ell,m}) = \kappa_k(z_{k,1}, \dots, z_{k,m})$, $\ell, k = 1, \dots, n$. This requires problem-specific analysis but tends to perform well as we will see in Section 5.

2.3. Justification of the preconditioner and sampling restrictions. With M_{i+1} of the form (S.d) for any Λ_i, the implicit method (PP) expands as

$$0 \in T_i \partial G(x^{i+1}) + T_i K^* y^{i+1} + (x^{i+1} - x^i) - \Phi_i^{-1} \Lambda_i^*(y^{i+1} - y^i), \tag{2.10a}$$

$$0 \in \Sigma_{i+1} \partial F^*(y^{i+1}) - \Sigma_{i+1} K x^{i+1} - \Psi_{i+1}^{-1} \Lambda_i (x^{i+1} - x^i) + (y^{i+1} - y^i). \tag{2.10b}$$

This can easily be rearranged as

$$x^{i+1} := (I + T_i \partial G)^{-1}(x^i + \Phi_i^{-1} \Lambda_i^*(y^{i+1} - y^i) - T_i K^* y^{i+1}), \tag{2.11a}$$

$$y^{i+1} := (I + \Sigma_{i+1} \partial F^*)^{-1}(y^i + \Psi_{i+1}^{-1} \Lambda_i (x^{i+1} - x^i) + \Sigma_{i+1} K x^{i+1}). \tag{2.11b}$$

This method is still not explicit as the primal and dual updates potentially depend on each other. We will now show how the removal of this dependency leads to our choice of Λ_i in (S.d).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Indeed, due to the compatible block-separable structures (S) and (GF), multiplying (2.10a) by P_j, for $j = 1, \dots, m$, and (2.10b) by Q_ℓ, for $\ell = 1, \dots, n$, (2.11) can be split blockwise as

$$x_j^{i+1} = (I + \chi_{S(i)}(j)\tau_{j,i}\partial G_j)^{-1}(x_j^i + P_j[\Phi_i^{-1}\Lambda_i^*(y^{i+1} - y^i) - T_i K^* y^{i+1}]),$$
$$y_\ell^{i+1} = (I + \chi_{V(i+1)}(\ell)\sigma_{\ell,i+1}\partial F_\ell^*)^{-1}(y_\ell^i + Q_\ell[\Psi_{i+1}^{-1}\Lambda_i(x^{i+1} - x^i) + \Sigma_{i+1}Kx^{i+1}]).$$

For $S(i)$ and $V(i+1)$ to have the intended meaning that only the corresponding blocks are updated, we need to choose Λ_i such that

$$x_j^{i+1} = x_j^i \quad (j \notin S(i)) \qquad \text{and} \qquad y_\ell^{i+1} = y_\ell^i \quad (\ell \notin V(i+1)). \tag{2.12}$$

Since $P_j T_i = 0$, for $j \notin S(i)$, and $Q_\ell \Sigma_{i+1} = 0$, for $\ell \notin V(i+1)$, this is to say that

$$P_j\Phi_i^{-1}\Lambda_i^*(y^{i+1} - y^i) = 0 \qquad (j \notin S(i)) \quad \text{and} \tag{2.13a}$$
$$Q_\ell\Psi_{i+1}^{-1}\Lambda_i(x^{i+1} - x^i) = 0 \qquad (\ell \notin V(i+1)). \tag{2.13b}$$

Taking $\Lambda_i = KT_i^*\Phi_i^*$ would allow computing x^{i+1} before y^{i+1} and to satisfy (2.13a). If we further had $KT_i^*\Phi_i^* = \omega_i\Psi_{i+1}\Sigma_{i+1}K$, then also (2.13b) would hold and (S.d) would reproduce M_{i+1} of (2.4). Symmetrically, $\Lambda_i = -\Sigma_{i+1}\Psi_{i+1}K$ would make y^{i+1} independent of x^{i+1}. Such conditions will, however, rarely be satisfiable unless, deterministically, $S(i) \equiv \{1, \dots, m\}$ and $V(i+1) \equiv \{1, \dots, n\}$. Nevertheless, motivated by this, we designate subsets of blocks of x^{i+1} and y^{i+1} to be updated independently of the other variable. These are the subsets $\mathring{S}(i)$ and $\mathring{V}(i+1)$ in (S.e) and (S.f). Then picking Λ_i as in (S.d) achieves our objective:

LEMMA 2.5 *Suppose Assumption 2.1 holds. Then* (2.12) *and* (2.13) *hold.*

Proof. We already know that (2.13) implies (2.12). Using (S.d), (2.13) can be rewritten as

$$P_j\Phi_i^{-1}[\Phi_i\mathring{T}_iK^* - K^*\mathring{\Sigma}_{i+1}^*\Psi_{i+1}^*](y^{i+1} - y^i) = 0 \qquad (j \notin S(i)) \quad \text{and}$$
$$Q_\ell\Psi_{i+1}^{-1}[K\mathring{T}_i^*\Phi_i^* - \Psi_{i+1}\mathring{\Sigma}_{i+1}K](x^{i+1} - x^i) = 0 \qquad (\ell \notin V(i+1)).$$

Clearly $P_j\mathring{T}_iK^* = 0$ for $j \notin S(i)$. Therefore, the first condition holds if $P_jK^*\mathring{\Sigma}_{i+1}^* = 0$ for $j \notin S(i)$. This is to say that $j \notin \mathcal{V}^{-1}(\mathring{V}(i+1))$, which is guaranteed by ($\mathcal{V}$.b). Likewise, the second condition holds if $Q_\ell K\mathring{T}_i^* = 0$ for $\ell \notin V(i+1)$, which is also guaranteed by ($\mathcal{V}$.b). ∎

2.4. Overall structure of the proposed method. Defining the operators $T_i^\perp := T_i - \mathring{T}_i$, and $\Sigma_{i+1}^\perp := \Sigma_{i+1} - \mathring{\Sigma}_{i+1}$, we can now rewrite (2.11) as

$$q^{i+1} := \Phi_i^{-1}K^*\mathring{\Sigma}_{i+1}^*\Psi_{i+1}^*(y^{i+1} - y^i) + T_i^\perp K^* y^{i+1}, \tag{2.14a}$$
$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - \mathring{T}_iK^*y^i - q^{i+1}), \tag{2.14b}$$
$$z^{i+1} := \Psi_{i+1}^{-1}K\mathring{T}_i^*\Phi_i^*(x^{i+1} - x^i) + \Sigma_{i+1}^\perp Kx^{i+1}, \tag{2.14c}$$
$$y^{i+1} := (I + \Sigma_{i+1}\partial F^*)^{-1}(y^i + \mathring{\Sigma}_{i+1}Kx^i + z^{i+1}). \tag{2.14d}$$

Due to the first part of ($\mathcal{V}$.a), $P_jq^{i+1} = 0$ and $Q_\ell z^{i+1} = 0$ for $j \in \mathring{S}(i)$ and $\ell \in \mathring{V}(i+1)$. The second part of ($\mathcal{V}$.a) implies

$$T_i^\perp K^* Q_\ell = 0 \quad \text{and} \quad \Sigma_{i+1}^\perp K P_j, \quad \text{for } \ell \in V(i+1) \setminus \mathring{V}(i+1) \text{ and } j \in S(i) \setminus \mathring{S}(i).$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The first part of (V.a) and (V.b) moreover imply $\Sigma_{i+1}^{\perp}\Theta_{i+1} = \Sigma_{i+1}\Theta_{i+1} = \Psi_{i+1}^{-1}K\mathring{T}_i^*\Phi_i^*$ and $T_i^{\perp}B_{i+1}^* = T_iB_{i+1}^* = \Phi_i^{-1}K^*\mathring{\Sigma}_{i+1}\Psi_{i+1}$ for

$$\Theta_i := \sum_{j\in\mathring{S}(i)}\sum_{\ell\in\mathcal{V}(j)}\theta_{\ell,j,i}Q_\ell KP_j \qquad \text{with} \quad \theta_{\ell,j,i+1} := \frac{\tau_{j,i}\phi_{j,i}}{\sigma_{\ell,i+1}\psi_{\ell,i+1}} \quad \text{and}$$

$$B_i := \sum_{\ell\in\mathring{V}(i+1)}\sum_{j\in\mathcal{V}^{-1}(\ell)}b_{\ell,j,i}Q_\ell KP_j \qquad \text{with} \quad b_{\ell,j,i+1} := \frac{\sigma_{\ell,i+1}\psi_{\ell,i+1}}{\tau_{j,i}\phi_{j,i}}.$$

Letting $\mathring{x}^{i+1} := \sum_{j\in\mathring{S}(i)}P_jx^{i+1}$ and $\mathring{y}^{i+1} := \sum_{\ell\in\mathring{V}(i+1)}Q_\ell x^{i+1}$ we therefore obtain

$$\begin{aligned} q^{i+1} &:= \Phi_i^{-1}K^*\mathring{\Sigma}_{i+1}^*\Psi_{i+1}^*(\mathring{y}^{i+1} - y^i) + T_i^{\perp}K^*\mathring{y}^{i+1} \\ &= T_i^{\perp}[B_{i+1}^*(\mathring{y}^{i+1} - y^i) + \mathring{y}^{i+1}], \end{aligned} \tag{2.15a}$$

$$\begin{aligned} z^{i+1} &:= \Psi_{i+1}^{-1}K\mathring{T}_i^*\Phi_i^*(\mathring{x}^{i+1} - x^i) + \Sigma_{i+1}^{\perp}K\mathring{x}^{i+1} \\ &= \Sigma_{i+1}^{\perp}[\Theta_{i+1}(\mathring{x}^{i+1} - x^i) + \mathring{x}^{i+1}]. \end{aligned} \tag{2.15b}$$

Introducing w^{i+1} and v^{i+1} such that $\Psi_{i+1}^{\perp}w^{i+1} = z^{i+1}$ and $\Phi_i^{\perp}v^{i+1} = q^{i+1}$ and using (GF), we can write the method given by (2.14) and (2.15) as

$$\mathring{x}^{i+1} := (I + \mathring{T}_i\partial G)^{-1}(x^i - \mathring{T}_iK^*y^i), \tag{2.16a}$$

$$\mathring{y}^{i+1} := (I + \mathring{\Sigma}_{i+1}\partial F^*)^{-1}(y^i + \mathring{\Sigma}_{i+1}Kx^i), \tag{2.16b}$$

$$w^{i+1} := \Theta_{i+1}(\mathring{x}^{i+1} - x^i) + \mathring{x}^{i+1}, \tag{2.16c}$$

$$v^{i+1} := B_{i+1}^*(\mathring{y}^{i+1} - y^i) + \mathring{y}^{i+1}, \tag{2.16d}$$

$$x^{i+1} := (I + T_i^{\perp}\partial G)^{-1}(\mathring{x}^{i+1} - T_i^{\perp}v^{i+1}), \tag{2.16e}$$

$$y^{i+1} := (I + \Sigma_{i+1}^{\perp}\partial F^*)^{-1}(\mathring{y}^{i+1} + \Sigma_{i+1}^{\perp}w^{i+1}). \tag{2.16f}$$

Due to (GF), these operations can further be split into blockwise operations with no dependencies on so far uncomputed blocks. We delay making this explicitly until we are ready to present our final Algorithms 1 and 2 towards the end of the theoretical part of the paper.

We conclude the present structural development by explicitly stating what we have proved in the preceding paragraphs:

LEMMA 2.6 *Suppose that Assumption 2.1 holds. Then* (2.16) *is equivalent to* (PP).

3. A basic convergence estimate. Now that we have established the overall structure of the proposed algorithms in (2.16), we need to develop rules for the step length and testing parameters that yield a convergent method. This will require, in particular, the positive semi-definiteness of $Z_{i+1}M_{i+1}$ as we recall from the discussion leading up to (2.7). Indeed, since in general, without any strong convexity, we can only obtain gap (and weak) convergence, we need to refine that argument.

In Sections 3.1–3.5, we will derive some quite technical conditions that the step length parameters, testing parameters, and block selection probabilities need to satisfy. From these basic estimates, we then develop explicit convergence rates in the next section. In the final Section 3.6, we will also discuss permissible sampling patterns.

3.1. A bound on ergodic duality gaps. Recall the basis of the testing technique (2.5). In the single-block case ($T_i = \tau_iI$, $\Sigma_{i+1} = \sigma_{i+1}I$, $\Phi_{i+1} = \phi_{i+1}I$, and $\Psi_{i+1} = \psi_{i+1}I$), instead of using $\widehat{u} \in H^{-1}(0)$ and the operator-relative monotonicity (2.6) to eliminate H,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

using the convexity of G and F^* we can also estimate

$$\begin{aligned}
&\langle H(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}W_{i+1}} \\
&\quad \geq \phi_i\tau_i[G(x^{i+1}) - G(\widehat{x})] + \psi_{i+1}\sigma_{i+1}[F^*(y^{i+1}) - F^*(\widehat{y})] \\
&\qquad + \phi_i\tau_i\langle K^*y^{i+1}, x^{i+1} - \widehat{x}\rangle - \psi_{i+1}\sigma_i\langle Kx^{i+1}, y^{i+1} - \widehat{y}\rangle \\
&\quad =: \widetilde{\mathcal{G}}_{i+1}.
\end{aligned} \tag{3.1}$$

With this, (2.7) can be improved to

$$\frac{1}{2}\|u^N - \widehat{u}\|^2_{Z_{N+1}M_{N+1}} + \sum_{i=0}^{N-1}\left(\widetilde{\mathcal{G}}_{i+1} + \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}\right) \leq \frac{1}{2}\|u^0 - \widehat{u}\|^2_{Z_1M_1}.$$

We would like to develop the "preliminary gaps" $\widetilde{\mathcal{G}}_{i+1}$ into a (Lagrangian) duality gap

$$\mathcal{G}(x, y) := \big(G(x) + \langle\widehat{y}, Kx\rangle - F^*(\widehat{y})\big) - \big(G(\widehat{x}) + \langle y, K\widehat{x}\rangle - F^*(y)\big).$$

The first obstacle we face are the differing factors in front of G and F^*. This suggests to impose $\phi_i\tau_i = \psi_{i+1}\sigma_{i+1}$. For the PDHGM, it, however, turns out that $\phi_i\tau_i = \psi_i\sigma_i$. After taking care of some technical details, this can be dealt with by an index realignment argument [35].

With multiple blocks, we can get a similar estimate as (3.1) with the factors $\phi_{j,i}\tau_{j,i}$ in front of G_j and $\psi_{\ell,i+1}\sigma_{\ell,i+1}$ in front of F^*_ℓ. To derive a gap estimate, the preceding discussion suggests to impose $T_i\Phi_i = \bar{\eta}_i I$ and $\Sigma_{i+1}\Psi_{i+1} = \bar{\eta}_i I$ or $\Sigma_i\Psi_i = \bar{\eta}_i I$ for some scalar $\bar{\eta}_i > 0$. This kind of *coupling* between the blocks will be one of the main restrictions that we are faced with in the development of our method. In the stochastic setting, it turns out [35] that we can relax the coupling slightly: do it in expectation. Correspondingly, we assume for some $\bar{\eta}_i > 0$ either

$$\mathbb{E}[T_i^*\Phi_i^*] = \bar{\eta}_i I \quad \text{and} \quad \mathbb{E}[\Psi_{i+1}\Sigma_{i+1}] = \bar{\eta}_i I \quad (i \geq 1) \quad \text{or} \tag{3.2a}$$

$$\mathbb{E}[T_i^*\Phi_i^*] = \bar{\eta}_i I \quad \text{and} \quad \mathbb{E}[\Psi_i\Sigma_i] = \bar{\eta}_i I \quad (i \geq 1). \tag{3.2b}$$

The second condition is an extension of what we saw the standard PDHGM to satisfy. The first condition, which is off-by-one compared to the second, will, however, be the only alternative that doubly-stochastic methods can satisfy.

A further difficulty with developing (3.1) into a gap estimate are the remaining terms involving K. Even after rearrangements we can only get an ergodic estimate [35]. To express such estimates, corresponding to the conditions (3.2a) and (3.2b), we introduce

$$\zeta_N := \sum_{i=0}^{N-1}\bar{\eta}_i \quad \text{and} \quad \zeta_{*,N} := \sum_{i=1}^{N-1}\bar{\eta}_i \tag{3.3}$$

and the ergodic sequences

$$\widetilde{x}_N := \zeta_N^{-1}\mathbb{E}\left[\sum_{i=0}^{N-1} T_i^*\Phi_i^* x^{i+1}\right], \qquad \widetilde{y}_N := \zeta_N^{-1}\mathbb{E}\left[\sum_{i=0}^{N-1}\Sigma_{i+1}^*\Psi_{i+1}^* y^{i+1}\right],$$

$$\widetilde{x}_{*,N} := \zeta_{*,N}^{-1}\mathbb{E}\left[\sum_{i=1}^{N-1} T_i^*\Phi_i^* x^{i+1}\right], \qquad \widetilde{y}_{*,N} := \zeta_{*,N}^{-1}\mathbb{E}\left[\sum_{i=1}^{N-1}\Sigma_i^*\Psi_i^* y^i\right].$$

The coupling conditions (3.2a) and (3.2b) then produce two different ergodic gaps, $\mathcal{G}(\widetilde{x}_N, \widetilde{y}_N)$ and $\mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N})$. We demonstrate this in the next theorem from [35]. It forms the basis for our work in the remaining sections. The fundamental arguments for the proof are those that led to (2.7), however, the gap estimate requires significant additional technical work.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

THEOREM 3.1 *Suppose Assumption 2.1 (main structural condition) holds with $Z_{i+1}M_{i+1}$ positive semi-definite. Write $\Gamma := \sum_{j=1}^m \gamma_j P_j$ for $\gamma_j \geq 0$ the factor of (strong) convexity of G_j. With $\widetilde{\Gamma} = \sum_{j=1}^m \widetilde{\gamma}_j P_j \in \mathcal{L}(X;X)$, assuming one of the following alternatives to hold, let*

$$\widetilde{g}_N := \begin{cases} 0, & 0 \leq \widetilde{\Gamma} \leq \Gamma, \\ \zeta_N \mathcal{G}(\widetilde{x}_N, \widetilde{y}_N), & 0 \leq \widetilde{\Gamma} \leq \Gamma/2, \ \text{(3.2a) holds}, \\ \zeta_{*,N} \mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N}), & 0 \leq \widetilde{\Gamma} \leq \Gamma/2, \ \text{(3.2b) holds}. \end{cases}$$

Also define

$$\Xi_{i+1}(\widetilde{\Gamma}) := \begin{bmatrix} 2T_i\widetilde{\Gamma} & 2T_iK^* \\ -2\Sigma_{i+1}K & 0 \end{bmatrix},$$
$$D_{i+1}(\widetilde{\Gamma}) := Z_{i+2}M_{i+2} - Z_{i+1}(\Xi_{i+1}(\widetilde{\Gamma}) + M_{i+1}).$$

Then the iterates $u^i = (x^i, y^i)$ of (PP) satisfy for any $\widehat{u} \in H^{-1}(0)$ the estimate

$$\begin{aligned} &\frac{1}{2}\mathbb{E}\big[\|u^N - \widehat{u}\|^2_{Z_N M_N}\big] + \widetilde{g}_N \\ &\qquad \leq \frac{1}{2}\|u^0 - \widehat{u}\|^2_{Z_1M_1} + \sum_{i=0}^{N-1} \frac{1}{2}\mathbb{E}\big[\|u^{i+1} - \widehat{u}\|^2_{D_{i+1}(\widetilde{\Gamma})} - \|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}\big]. \end{aligned} \tag{3.4}$$

Proof. This is [35, Theorem 5.5] with

$$\Delta_{i+1}(\widetilde{\Gamma}) := \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{D_{i+1}(\widetilde{\Gamma})} - \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}$$

and the condition $\widetilde{\Gamma} = \Gamma$ relaxed to $0 \leq \widetilde{\Gamma} \leq \Gamma$, which is possible because if g_j is strongly convex with factor $\gamma_j > 0$, then it is strongly convex with any smaller non-negative factor. Moreover, [35, Example 5.1] shows that the blockwise structure (GF), (S) has an ergodic convexity property that produces the gaps $\mathcal{G}(\widetilde{x}_N, \widetilde{y}_N)$ and $\mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N})$. ∎

In the standard PDHGM we can ensure that $D_{i+1}(\widetilde{\Gamma}) \simeq 0$ [35]. However, in our present setting, we will not generally be able to enforce this, so these operators will introduce a penalty in (3.4). A lot of our remaining work will consist of controlling this penalty. We also need to estimate from below and show that $Z_N M_N$ is positive semi-definite.

3.2. Notations and assumptions. For convenience, we introduce

$$\begin{aligned} \hat{\tau}_{j,i} &:= \tau_{j,i}\chi_{S(i)}(j), & \hat{\sigma}_{\ell,i} &:= \sigma_{\ell,i}\chi_{V(i)}(\ell), \\ \pi_{j,i} &:= \mathbb{P}[j \in S(i) \mid \mathcal{O}_{i-1}], & \nu_{\ell,i+1} &:= \mathbb{P}[\ell \in V(i+1) \mid \mathcal{O}_{i-1}], \\ \mathring{\pi}_{j,i} &:= \mathbb{P}[j \in \mathring{S}(i) \mid \mathcal{O}_{i-1}], & \mathring{\nu}_{\ell,i+1} &:= \mathbb{P}[\ell \in \mathring{V}(i+1) \mid \mathcal{O}_{i-1}]. \end{aligned}$$

The first two denote "effective" step lengths at iteration i, while the rest is shorthand for the probabilities of the primal block j or the dual block ℓ being contained in the corresponding set at iteration i. Recalling (S.d), we also write

$$\Lambda_i = \sum_{j=1}^m \sum_{\ell \in \mathcal{V}(j)} \lambda_{\ell,j,i} Q_\ell K P_j \qquad \text{with} \tag{3.5}$$
$$\lambda_{\ell,j,i} := \phi_{j,i}\hat{\tau}_{j,i}\chi_{\mathring{S}(i)}(j) - \psi_{\ell,i+1}\hat{\sigma}_{\ell,i+1}\chi_{\mathring{V}(i+1)}(\ell).$$

We require the following technical assumption, which we will verify through explicit step length and a testing parameter update rules development in the next section. We indicate the rough intended use of each condition in parentheses after the statement.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

ASSUMPTION 3.2 (step length parameter restrictions). We assume for each $i \in \mathbb{N}$ the following, with constants independent of i, and the *same alternatives* holding for each i:

(a) We are given $(\kappa_1, \ldots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$ (see Definition 2.2), and for some $\delta \in (0,1)$,

$$(1-\delta)\psi_{\ell,i+1} \geq \kappa_\ell(\lambda_{\ell,1,i}^2 \phi_{1,i}^{-1}, \ldots, \lambda_{\ell,m,i}^2 \phi_{m,i}^{-1}) \qquad (\ell = 1, \ldots, n).$$

(This condition generalises the condition $\tau\sigma\|K\|^2 < 1$ for the standard PDHGM, needed to ensure the positivity of the local metric $Z_{i+1}M_{i+1}$.)

(b) We are given $\eta_i \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$ and $\eta_{\tau,i}^\perp, \eta_{\sigma,i}^\perp \in \mathcal{R}(\mathcal{O}_{i-1}; [0,\infty))$ such that

$$\eta_{i+1} \geq \eta_i, \quad \eta_i \cdot \min_j(\pi_{j,i} - \mathring{\pi}_{j,i}) \geq \eta_{\tau,i}^\perp, \quad \text{and} \quad \eta_{i+1} \cdot \min_\ell(\nu_{\ell,i+1} - \mathring{\nu}_{\ell,i+1}) \geq \eta_{\sigma,i}^\perp.$$

(This is needed to annihilate the off-diagonal of D_{i+1} in the penalty term.)

(c) Either

$$\text{(c-i)} \quad \mathbb{E}[\eta_{\tau,i}^\perp - \eta_{\sigma,i}^\perp] = \text{constant} \qquad \text{or} \qquad \text{(c-ii)} \quad \eta_{\tau,i}^\perp = 0 \text{ and } \eta_{\sigma,i}^\perp = \eta_{i+1}.$$

(These are needed to ensure the coupling conditions (3.2a) or (3.2b), respectively.)

(d) The step lengths parameters satisfy

$$\text{(3.6a)} \qquad \tau_{j,i} = \begin{cases} \frac{\eta_i - \phi_{j,i-1}\tau_{j,i-1}\chi_{S(i-1)\setminus\mathring{S}(i-1)}(j)}{\phi_{j,i}\mathring{\pi}_{j,i}}, & j \in \mathring{S}(i), \\ \frac{\eta_{\tau,i}^\perp}{\phi_{j,i}(\pi_{j,i} - \mathring{\pi}_{j,i})}, & j \in S(i) \setminus \mathring{S}(i), \end{cases}$$

$$\text{(3.6b)} \qquad \sigma_{j,i+1} = \begin{cases} \frac{\eta_i - \psi_{j,i}\sigma_{j,i}\chi_{V(i)\setminus\mathring{V}(i)}(j)}{\psi_{j,i+1}\mathring{\nu}_{\ell,i+1}}, & j \in \mathring{V}(i+1), \\ \frac{\eta_{\sigma,i}^\perp}{\psi_{j,i+1}(\nu_{\ell,i+1} - \mathring{\nu}_{\ell,i+1})}, & j \in V(i+1) \setminus \mathring{V}(i+1). \end{cases}$$

For $i = 0$ we take $\tau_{j,-1} := 0$ and $\sigma_{j,0} := 0$.
(This rule is also needed to annihilate the off-diagonal of D_{i+1} in the penalty term.)

(e) Let $\gamma_j \geq 0$ be the factor of (strong) convexity of G_j and $\widetilde{\gamma}_j \in [0, \gamma_j]$, $j = 1, \ldots, m$. Also let $\alpha_i > 0$ and define

$$\text{(3.7a)} \quad q_{j,i+2}(\widetilde{\gamma}_j) := \Big(\mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|\mathcal{O}_i] + \alpha_i|\mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|\mathcal{O}_i]| - \delta\phi_{j,i}\Big)\chi_{S(i)}(j),$$

$$\text{(3.7b)} \quad h_{j,i+2}(\widetilde{\gamma}_j) := \mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|\mathcal{O}_{i-1}] + \alpha_i^{-1}|\mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|\mathcal{O}_i]|.$$

Then for some $C_x > 0$ either

$$\text{(3.8a)} \quad \|x_j^{i+1} - \widehat{x}_j\|^2 \leq C_x \qquad (j = 1, \ldots, m) \quad \text{or}$$

$$\text{(3.8b)} \quad h_{j,i+2}(\widetilde{\gamma}_j) \leq 0 \quad \text{and} \quad q_{j,i+2}(\widetilde{\gamma}_j) \leq 0 \qquad (j = 1, \ldots, m).$$

(This is needed to bound the primal components in the penalty term.)

(f) For some $C_y > 0$ either

$$\text{(3.9a)} \quad \mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}|\mathcal{O}_i] \geq 0, \quad \|y_\ell^{i+1} - \widehat{y}_\ell\|^2 \leq C_y \quad (\ell = 1, \ldots, n) \quad \text{or}$$

$$\text{(3.9b)} \quad \mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}|\mathcal{O}_i] = 0 \qquad (\ell = 1, \ldots, n).$$

(This is needed to bound the dual components in the penalty term.)

ETNA
Kent State University and
Johann Radon Institute (RICAM)

It is important that Assumption 3.2 is consistent with Assumption 2.1, in particular that the step lengths generated by the former are non-negative. We will prove this in Lemma 3.5. Before this, we state the main goal of the present section, the following specialisation of Theorem 3.1.

PROPOSITION 3.3 *Suppose Assumption 2.1 (main structural condition) and Assumption 3.2 (step length restrictions) hold. Then the iterates of* (PP) *satisfy for any* $\widehat{u} \in H^{-1}(0)$ *the estimate*

$$\sum_{j=1}^{m} \frac{\delta}{2\mathbb{E}[\phi_{j,N}^{-1}]} \cdot \mathbb{E}\big[\|x_j^N - \widehat{x}_j\|\big]^2 + \widetilde{g}_N \leq \frac{1}{2}\|u^0 - \widehat{u}\|_{Z_0 M_0}^2 + \sum_{j=1}^{m} \frac{1}{2} d_{j,N}^x(\widetilde{\gamma}_j) + \sum_{\ell=1}^{n} \frac{1}{2} d_{\ell,N}^y, \tag{3.10}$$

where

$$\widetilde{g}_N := \begin{cases} \zeta_N \mathcal{G}(\widetilde{x}_N, \widetilde{y}_N), & \text{case (c-i) and } \widetilde{\gamma}_j \leq \gamma_j/2 \text{ for all } j, \\ \zeta_{*,N} \mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N}), & \text{case (c-ii) and } \widetilde{\gamma}_j \leq \gamma_j/2 \text{ for all } j, \\ 0, & \text{otherwise}, \end{cases} \tag{3.11a}$$

$$d_{j,N}^x(\widetilde{\gamma}_j) := \sum_{i=0}^{N-1} \delta_{j,i+2}^x(\widetilde{\gamma}_j), \qquad d_{\ell,N}^y := \sum_{i=0}^{N-1} \delta_{\ell,i+2}^y, \tag{3.11b}$$

$$\delta_{j,i+2}^x(\widetilde{\gamma}_j) := 4C_x \mathbb{E}[\max\{0, q_{j,i+2}(\widetilde{\gamma}_j)\}] + C_x \mathbb{E}[\max\{0, h_{j,i+2}(\widetilde{\gamma}_j)\}], \quad \text{and} \tag{3.11c}$$

$$\delta_{\ell,i+2}^y := 9C_y \mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}]. \tag{3.11d}$$

Proof. We use Lemma 3.9 or Lemma 3.10 (see below) to verify one of the coupling conditions (3.2a) or (3.2b). Then we obtain (3.4) from Theorem 3.1. Next, we use Lemmas 3.4 and 3.7 (see below) to estimate $Z_{N+1}M_{N+1} \geq \left(\begin{smallmatrix} \delta\Phi_N & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and

$$\mathbb{E}\big[\|u^{i+1} - \widehat{u}\|_{D_{i+1}(\widetilde{\Gamma})}^2 - \|u^{i+1} - u^i\|_{Z_{i+1}M_{i+1}}^2\big] \leq \sum_{j=1}^{m} \delta_{j,i+2}^x(\widetilde{\gamma}_j) + \sum_{\ell=1}^{n} \delta_{\ell,i+2}^y.$$

Therefore (3.4) yields

$$\frac{\delta}{2}\mathbb{E}\big[\|x^N - \widehat{x}\|_{\Phi_N}^2\big] + \widetilde{g}_N \leq \frac{1}{2}\|u^0 - \widehat{u}\|_{Z_0 M_0}^2 + \frac{1}{2}\sum_{i=0}^{N-1}\left(\sum_{j=1}^{m} \delta_{j,i+2}^x(\widetilde{\gamma}_j) + \sum_{\ell=1}^{n} \delta_{\ell,i+2}^y\right).$$

By Hölder's inequality it follows that

$$\mathbb{E}\big[\|x^N - \widehat{x}\|_{\Phi_N}^2\big] = \sum_{j=1}^{m} \mathbb{E}\big[\phi_{j,N}\|x_j^N - \widehat{x}_j\|^2\big] \geq \sum_{j=1}^{m} \mathbb{E}\big[\|x_j^N - \widehat{x}_j\|\big]^2 / \mathbb{E}[\phi_{j,N}^{-1}].$$

The estimate (3.10) is now immediate. ∎

3.3. A lower bound on the local metric.

LEMMA 3.4 *Suppose that Assumption 2.1 (main structural condition) and Assumption 3.2(a) hold. Then* $Z_{i+1}M_{i+1} \geq \left(\begin{smallmatrix} \delta\Phi_i & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

Proof. Since Φ_{i+1} is self-adjoint and positive definite, using (2.9) and Cauchy's inequality, for any $\delta \in (0,1)$, we deduce

$$Z_{i+1}M_{i+1} = \begin{bmatrix} \Phi_i & -\Lambda_i^* \\ -\Lambda_i & \Psi_{i+1} \end{bmatrix} \geq \begin{bmatrix} \delta\Phi_i & 0 \\ 0 & \Psi_{i+1} - \frac{1}{1-\delta}\Lambda_i \Phi_i^{-1} \Lambda_i^* \end{bmatrix}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

We therefore require $(1-\delta)\Psi_{i+1} \geq \Lambda_i \Phi_i^{-1} \Lambda_i^*$, which can be expanded as

$$(1-\delta)\sum_{\ell=1}^{n} \psi_{\ell,i+1} Q_\ell \geq \sum_{j=1}^{m}\sum_{\ell,k=1}^{n} \lambda_{\ell,j,i}\lambda_{k,j,i}\phi_{j,i}^{-1} Q_\ell K P_j K^* Q_k.$$

This follows from Definition 2.2(i) with $z_{\ell,j} := \lambda_{\ell,j,i}^2 \phi_{j,i}^{-1}$. ∎

3.4. Bounds on the penalty terms. The structural setup (S) gives

$$D_{i+1}(\widetilde{\Gamma}) = \begin{bmatrix} \Phi_{i+1} - \Phi_i(I + 2T_i\widetilde{\Gamma}) & \Lambda_i^* - \Lambda_{i+1}^* - 2\Phi_i T_i K^* \\ 2\Psi_{i+1}\Sigma_{i+1}K + \Lambda_i - \Lambda_{i+1} & \Psi_{i+2} - \Psi_{i+1} \end{bmatrix} \tag{3.12}$$
$$\simeq \begin{bmatrix} \Phi_{i+1} - \Phi_i(I + 2T_i\widetilde{\Gamma}) & A_{i+2}^* \\ A_{i+2} & \Psi_{i+2} - \Psi_{i+1} \end{bmatrix} \quad \text{for}$$
$$A_{i+2} := (\Psi_{i+1}\Sigma_{i+1}K - \Lambda_{i+1}) + (\Lambda_i - KT_i^*\Phi_i^*).$$

LEMMA 3.5 *Suppose that Assumption 2.1 (main structural condition) and Assumptions 3.2(b) and 3.2(d) hold. Then*

$$\mathbb{E}[A_{i+2}|\mathcal{O}_i](x^{i+1} - x^i) = 0, \quad \mathbb{E}[A_{i+2}^*|\mathcal{O}_i](y^{i+1} - y^i) = 0, \quad \mathbb{E}[A_{i+2}^*|\mathcal{O}_{i-1}] = 0. \tag{3.13}$$

Moreover, if $\phi_{j,i}, \psi_{\ell,i+1} > 0$ for all $\in \mathbb{N}$, then $\tau_{j,i}, \sigma_{\ell,i+1} \geq 0$ for all $i \in \mathbb{N}$. In particular, Assumption 3.2 is consistent with Assumption 2.1 requiring $\tau_{j,i}, \sigma_{\ell,i+1} \geq 0$ and $\phi_{j,i}, \psi_{\ell,i+1} > 0$ for all $i \in \mathbb{N}$, $j = 1, \ldots, m$ and $\ell = 1, \ldots, n$.

Proof. We start by claiming that

$$\mathbb{E}[\lambda_{\ell,j,i+1}|\mathcal{O}_i] = \psi_{\ell,i+1}\hat{\sigma}_{\ell,i+1}(1 - \chi_{\mathring{V}(i+1)}(\ell)) - \phi_{j,i}\hat{\tau}_{j,i}(1 - \chi_{\mathring{S}(i)}(j)) \tag{3.14}$$

whenever $\ell \in \mathcal{V}(j)$. Indeed, inserting (3.5) into (3.14), we see the former to be satisfied if (for any given η_{i+1})

$$\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}\chi_{\mathring{S}(i+1)}(j)|\mathcal{O}_i] = \eta_{i+1} - \phi_{j,i}\hat{\tau}_{j,i}(1 - \chi_{\mathring{S}(i)}(j)) \geq 0 \quad \text{and} \tag{3.15a}$$
$$\mathbb{E}[\psi_{\ell,i+2}\hat{\sigma}_{\ell,i+2}\chi_{\mathring{V}(i+2)}(\ell)|\mathcal{O}_i] = \eta_{i+1} - \psi_{\ell,i+1}\hat{\sigma}_{\ell,i+1}(1 - \chi_{\mathring{V}(i+1)}(\ell)) \geq 0 \tag{3.15b}$$

over $j = 1, \ldots, m$, $\ell = 1, \ldots, n$, and $i \geq -1$, taking $\mathring{S}(-1) = \{1, \ldots, m\}$ and $\mathring{V}(0) = \{1, \ldots, n\}$.

We can also write (3.15a) as

$$\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}\chi_{\mathring{S}(i+1)}(j)|\mathcal{O}_i] = \eta_{i+1} - \phi_{j,i}\tau_{j,i}\chi_{S(i)\setminus\mathring{S}(i)}(j) \geq 0. \tag{3.16}$$

If $j \notin S(i) \setminus \mathring{S}(i)$, since $\eta_{i+1} \geq 0$ by Assumption 3.2(b), it is clear that the inequality in (3.16) holds. If $j \in S(i) \setminus \mathring{S}(i)$, using the corresponding case of (3.6a), we rewrite the inequality as $\eta_{i+1} \geq \eta_{\tau,i}^{\perp}/(\pi_{j,i} - \mathring{\pi}_{j,i})$. This is verified by Assumption 3.2(b). Comparing to Assumption 3.2(d), the inequality in (3.16) now inductively verifies, as claimed, $\tau_{j,i+1} \geq 0$ for all $i \in \mathbb{N}$ provided that $\phi_{j,i} > 0$ for all $i \in \mathbb{N}$.

To verify the equality in (3.16), let $\mathcal{O}_i^+ \supset \mathcal{O}_i$ be the smallest σ-algebra also containing the set $\{\omega \in \Omega \mid j \in \mathring{S}(\omega)(i+1)\}$ (now not abusing notation for random variables, with ω standing for the random realisation that we typically omit). By Assumption 3.2(d), more precisely (3.6b) shifted from i to $i+1$, we see that $\phi_{j,i+1}\tau_{j,i+1}$ is $\mathcal{O}_i^+$-measurable. Therefore,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

by standard properties of conditional expectations (see, e.g., [32])

$$\begin{aligned}\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}\chi_{\mathring{S}(i+1)}(j)|\mathcal{O}_i] &= \mathbb{E}[\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}\chi_{\mathring{S}(i+1)}(j)|\mathcal{O}_i^+]|\mathcal{O}_i]\\ &= \mathbb{E}[\mathbb{E}[\phi_{j,i+1}\tau_{j,i+1}|\mathcal{O}_i^+]|\mathcal{O}_i]\\ &= \mathbb{E}[\mathbb{E}[1|\mathcal{O}_i^+]\phi_{j,i+1}\tau_{j,i+1}|\mathcal{O}_i]\\ &= \mathbb{E}[\mathring{\pi}_{j,i+1}\phi_{j,i+1}\tau_{j,i+1}|\mathcal{O}_i].\end{aligned} \tag{3.17}$$

Further expanding with (3.6b) shifted from i to $i+1$ and using $\eta_{i+1} \in \mathcal{R}(\mathcal{O}_i;(0,\infty))$ from Assumption 3.2(b), we obtain

$$\begin{aligned}\mathbb{E}[\mathring{\pi}_{j,i+1}\phi_{j,i+1}\tau_{j,i+1}|\mathcal{O}_i] &= \mathbb{E}[\eta_{i+1} - \phi_{j,i}\tau_{j,i}\chi_{S(i)\setminus\mathring{S}(i)}(j)|\mathcal{O}_i]\\ &= \eta_{i+1} - \phi_{j,i}\tau_{j,i}\chi_{S(i)\setminus\mathring{S}(i)}(j).\end{aligned} \tag{3.18}$$

This verifies the equality in (3.16). Thus (3.15a) holds. Similarly we can verify (3.15b) and $\sigma_{\ell,i+1} \geq 0$. Thus (3.14) holds, as does the non-negativity claim on the dual step lengths.

Using ($\mathcal{V}$.b) and (3.14), we now observe that $\lambda_{\ell,j,i}$ satisfies

$$\lambda_{\ell,j,i} = 0 \qquad (j \notin S(i) \text{ or } \ell \notin V(i+1)) \quad \text{and} \tag{3.19a}$$

$$\mathbb{E}[\lambda_{\ell,j,i+1}|\mathcal{O}_i] = \widetilde{\lambda}_{\ell,j,i+1} \qquad (j = 1,\dots,m;\ \ell \in \mathcal{V}(j)), \tag{3.19b}$$

for $\widetilde{\lambda}_{\ell,j,i+1} := \psi_{\ell,i+1}\hat{\sigma}_{\ell,i+1} + \lambda_{\ell,j,i} - \phi_{j,i}\hat{\tau}_{j,i}$. Using (2.12), which follows from Lemma 2.5, (3.13) expands as

$$\mathbb{E}[\lambda_{\ell,j,i+1}|\mathcal{O}_i] = \widetilde{\lambda}_{\ell,j,i+1} \qquad (j \in S(i),\ \ell \in \mathcal{V}(j)), \tag{3.20a}$$

$$\mathbb{E}[\lambda_{\ell,j,i+1}|\mathcal{O}_i] = \widetilde{\lambda}_{\ell,j,i+1} \qquad (\ell \in V(i+1),\ j \in \mathcal{V}^{-1}(\ell)), \quad \text{and} \tag{3.20b}$$

$$\mathbb{E}[\lambda_{\ell,j,i+1}|\mathcal{O}_{i-1}] = \mathbb{E}[\widetilde{\lambda}_{\ell,j,i+1}|\mathcal{O}_{i-1}] \qquad (j = 1,\dots,m;\ \ell \in \mathcal{V}(j)). \tag{3.20c}$$

Clearly (3.19b) implies (3.20a) and (3.20b). Moreover, applying $\mathbb{E}[\,\cdot\,|\mathcal{O}_{i-1}]$ to (3.19b) and using standard properties of nested conditional expectations we obtain (3.20c). We have therefore verified (3.13). ∎

COROLLARY 3.6 *Suppose that Assumptions 3.2(b) and 3.2(d) hold. Then*

$$\begin{aligned}\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}|\mathcal{O}_i] &= \eta_{i+1} + \eta^{\perp}_{\tau,i+1} - \eta^{\perp}_{\tau,i} \quad \textit{and}\\ \mathbb{E}[\psi_{\ell,i+2}\hat{\sigma}_{\ell,i+2}|\mathcal{O}_i] &= \eta_{i+1} + \eta^{\perp}_{\sigma,i+1} - \eta^{\perp}_{\sigma,i}.\end{aligned}$$

Proof. Arguing analogously to (3.17) and (3.18) with the cases $j \in S(i) \setminus \mathring{S}(i)$ and $\ell \in V(i+1) \setminus \mathring{V}(i+1)$ of Assumption 3.2(d), we deduce that

$$\begin{aligned}\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}(1 - \chi_{\mathring{S}(i+1)}(j))|\mathcal{O}_i] &= \eta^{\perp}_{\tau,i+1} \quad \text{and}\\ \mathbb{E}[\psi_{\ell,i+2}\hat{\sigma}_{\ell,i+2}(1 - \chi_{\mathring{V}(i+2)}(\ell))|\mathcal{O}_i] &= \eta^{\perp}_{\sigma,i+1}.\end{aligned}$$

Combined with (3.15) (in the proof of Lemma 3.5) these imply the claim. ∎

For the next lemma we recall the coordinate notation x_j and y_ℓ from (2.8).

LEMMA 3.7 *Suppose Assumption 2.1 (main structural condition) and Assumption 3.2 (step length parameter restrictions) hold. Then*

$$\mathbb{E}[\|u^{i+1} - \widehat{u}\|^2_{D_{i+1}(\widetilde{\Gamma})} - \|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}] \leq \sum_{j=1}^{m} \delta^x_{j,i+2}(\widetilde{\gamma}_j) + \sum_{\ell=1}^{n} \delta^y_{\ell,i+2},$$

where $\delta^x_{j,i+2}(\widetilde{\gamma}_j)$ and $\delta^y_{\ell,i+2}$ are given in (3.11c) and (3.11d), respectively.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Proof. Since $u^{i+1} \in \mathcal{R}(\mathcal{O}_i; X \times Y)$ and $u^i \in \mathcal{R}(\mathcal{O}_{i-1}; X \times Y)$, standard nesting properties of conditional expectations show

$$\begin{aligned}\mathbb{E}[\|u^{i+1} - \widehat{u}\|^2_{D_{i+1}(\widetilde{\Gamma})}] = \mathbb{E}\big[&\|u^{i+1} - u^i\|^2_{\mathbb{E}[D_{i+1}(\widetilde{\Gamma})|\mathcal{O}_i]} + \|u^i - \widehat{u}\|^2_{\mathbb{E}[D_{i+1}(\widetilde{\Gamma})|\mathcal{O}_{i-1}]} \\ &+ 2\langle u^{i+1} - u^i, u^i - \widehat{u}\rangle_{\mathbb{E}[D_{i+1}(\widetilde{\Gamma})|\mathcal{O}_i]}\big].\end{aligned} \tag{3.21}$$

By Lemma 3.5, (3.13) holds. Using (3.12), we therefore expand (3.21) into

$$\begin{aligned}\mathbb{E}[\|u^{i+1} - \widehat{u}\|^2_{D_{i+1}(\widetilde{\Gamma})}] = \mathbb{E}\big[&\|x^{i+1} - x^i\|^2_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|\mathcal{O}_i]} \\ &+ \|x^i - \widehat{x}\|^2_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|\mathcal{O}_{i-1}]} \\ &+ \|y^{i+1} - y^i\|^2_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_i]} + \|y^i - \widehat{y}\|^2_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_{i-1}]} \\ &+ 2\langle x^{i+1} - x^i, x^i - \widehat{x}\rangle_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|\mathcal{O}_i]} \\ &+ 2\langle y^{i+1} - y^i, y^i - \widehat{y}\rangle_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_i]}\big].\end{aligned}$$

By Assumption 3.2(f), $\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_i] \geq 0$. Standard properties of conditional expectations guarantee that $\mathbb{E}[\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_i]|\mathcal{O}_{i-1}] = \mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_{i-1}]$. By Lemma 3.4, moreover, it holds that

$$-\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} \leq -\delta\|x^{i+1} - x^i\|_{\Phi_i}.$$

The use of Cauchy's inequality for arbitrary factors $\alpha_i, \beta_i > 0$ therefore yields

$$\begin{aligned}&\mathbb{E}\big[\|u^{i+1} - \widehat{u}\|^2_{D_{i+1}(\widetilde{\Gamma})} - \|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}\big] \\ &\quad= \mathbb{E}\big[\|x^{i+1} - x^i\|^2_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|\mathcal{O}_i] + \alpha_i|\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|\mathcal{O}_i]| - \delta\Phi_i} \\ &\qquad+ \|x^i - \widehat{x}\|^2_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|\mathcal{O}_{i-1}] + \alpha_i^{-1}|\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|\mathcal{O}_i]|} \\ &\qquad+ (1+\beta_i)\|y^{i+1} - y^i\|^2_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_i]} \\ &\qquad+ (1+\beta_i^{-1})\|y^i - \widehat{y}\|^2_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|\mathcal{O}_{i-1}]}\big].\end{aligned}$$

Here we write $|\sum_{j=1}^m c_j P_j| := \sum_{j=1}^m |c_j| P_j$. Therefore, by choosing $\beta_i = 1/2$, splitting the estimates into blocks, and using Assumptions 3.2(e) and 3.2(f), we obtain the claim. ∎

It is relatively easy to satisfy Assumption 3.2(f) and to bound $\delta^y_{\ell,i+2}$. To estimate $\delta^x_{j,i+2}(\widetilde{\gamma}_j)$, we need to derive more involved update rules. We next construct one example.

EXAMPLE 3.8 (Random primal test updates). If (3.8a) holds, then take $\rho_j \geq 0$, otherwise take $\rho_j = 0$ $(j = 1, \ldots, m)$. Set

$$\phi_{j,i+1} := \phi_{j,i}(1 + 2\widetilde{\gamma}_j\hat{\tau}_{j,i}) + 2\rho_j\pi_{j,i}^{-1}\chi_{S(i)}(j) \qquad (j = 1, \ldots, m; i \in \mathbb{N}). \tag{3.22}$$

Then it is not difficult to show that $\phi_{j,i+1} \in \mathcal{R}(\mathcal{O}_i; (0, \infty))$ and $\delta^x_{j,i+2}(\widetilde{\gamma}_j) = 18C_x\rho_j$.

If we set $\rho_j = 0$ and have just a single deterministically updated block, then (3.22) is the standard rule (2.1) with $\phi_i = \tau_i^{-2}$. The role of $\rho_j > 0$ is to ensure some (slower) acceleration on non-strongly-convex blocks with $\widetilde{\gamma}_j = 0$. This is necessary for convergence rate estimates.

The difficulty with (3.22) is that the coupling parameter η_{i+1} will depend on the random realisations of $S(i)$ through $\phi_{j,i+1}$. This will require communication in a parallel implementation of the algorithm. We therefore desire to update $\phi_{j,i+1}$ deterministically. We delay the introduction of an appropriate update rule to Section 4.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

3.5. Satisfying the coupling conditions. We still need to satisfy either one of the coupling conditions (3.2) to obtain gap estimates.

LEMMA 3.9 *Suppose that Assumption 2.1 (main structural condition), Assumptions 3.2(d), 3.2(b), and (c-i) hold. Then the coupling condition* (3.2a) *holds.*

Proof. The condition (3.2a) holds if $\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}] = \bar{\eta}_{i+1} = \mathbb{E}[\psi_{\ell,i+2}\hat{\sigma}_{\ell,i+2}]$ for some $\bar{\eta}_{i+1}$ for all $j = 1,\dots,m$ and $\ell = 1,\dots,n$. Taking $\bar{\eta}_{i+1} := \mathbb{E}[\eta_{i+1} + \eta^{\perp}_{\tau,i+1} - \eta^{\perp}_{\tau,i}]$, the claim follows from Corollary 3.6 and Assumption 3.2 (c-i). ∎

The alternative coupling condition (3.2b) requires

$$\mathbb{E}[\phi_{j,i+1}\hat{\tau}_{j,i+1}] = \bar{\eta}_{i+1} = \mathbb{E}[\psi_{\ell,i+1}\hat{\sigma}_{\ell,i+1}]$$

for some $\bar{\eta}_{i+1}$. By Corollary 3.6, this holds when

$$\mathbb{E}[\eta_{i+1} + \eta^{\perp}_{\tau,i+1} - \eta^{\perp}_{\tau,i}] = \bar{\eta}_i = \mathbb{E}[\eta_i + \eta^{\perp}_{\sigma,i} - \eta^{\perp}_{\sigma,i-1}]. \tag{3.23}$$

It is not clear how to satisfy this simultaneously with Assumption 3.2 (c-i), so we use (c-ii).

LEMMA 3.10 *Suppose that Assumption 2.1 (main structural condition), Assumptions 3.2(d) and (c-ii) hold. Then Assumption 3.2(b) holds if and only if* $\mathring{V}(i+1) = \emptyset$ *and* $V(i+1) = \{1,\dots,n\}$. *When this is the case, the coupling condition* (3.2b) *holds, necessarily* $S(i) = \mathring{S}(i)$, *and*

$$\tau_{j,i} = \eta_i/(\phi_{j,i}\mathring{\pi}_{j,i}) \qquad (j \in S(i)), \tag{3.24a}$$

$$\sigma_{j,i+1} = \eta_{i+1}/\psi_{j,i+1} \qquad (j \in \mathcal{V}(S(i))). \tag{3.24b}$$

Proof. Assumption 3.2 (c-ii), i.e., $\eta^{\perp}_{\tau,i} = 0$ and $\eta^{\perp}_{\sigma,i} = \eta_{i+1}$ reduces Assumption 3.2(b) to $\min_\ell(\nu_{\ell,i+1} - \mathring{\nu}_{\ell,i+1}) \geq 1$. This holds if and only hold if $\nu_{\ell,i+1} \equiv 1$ and $\mathring{\nu}_{\ell,i+1} \equiv 0$ for all $\ell = 1,\dots,m$. This holds, as claimed, if and only if $\mathring{V}(i+1) = \emptyset$, $V(i+1) = \{1,\dots,n\}$. Clearly in this case ($\mathcal{V}$) holds if and only if $S(i) = \mathring{S}(i)$. With Assumption 3.2(b) verified, Corollary 3.6 now rewrites (3.2b) as (3.23). This is clearly verified by $\eta^{\perp}_{\tau,i} = 0$ and $\eta^{\perp}_{\sigma,i} = \eta_{i+1}$. Finally, (3.24) is a specialisation of Assumption 3.2(d) to the choices of (c-ii). ∎

REMARK 3.11 We had to impose full dual updates to satisfy (3.2b). This is akin to most existing primal-dual coordinate descent methods [3, 15, 33]. The algorithms in [24, 26, 40] are more closely related to our method, however, only [40] provides convergence rates for single-block sampling schemes under full strong convexity of both G and F^*.

3.6. Sampling patterns. There are not many possible fully deterministic sampling patterns allowed by ($\mathcal{V}$) with Assumption 3.2. Indeed, (3.15a) reads in the deterministic setting

$$\phi_{j,i+1}\tau_{j,i+1}\chi_{\mathring{S}(i+1)}(j) + \phi_{j,i}\hat{\tau}_{j,i}\chi_{S(i)\setminus\mathring{S}(i)}(j)) = \eta_{i+1}.$$

Since $\eta_{i+1} > 0$, $j \notin S(i) \setminus \mathring{S}(i)$ implies $j \in \mathring{S}(i+1)$, which implies $j \notin S(i+i) \setminus \mathring{S}(i+1)$. Therefore, once in the independently updated set, the block j will always stay there. Due to ($\mathcal{V}$.b), if $\mathring{V}(i+1) \neq \emptyset$ consistently, for most K, the set $S(i)$ will grow. Therefore, after a small number of iterations N, either $j \in \mathring{S}(i)$, for $i \geq N$, or $j \in S(i) = \{1,\dots,n\}$. Similar considerations hold for the dual blocks. Therefore, the way each block is updated in deterministic methods is, after a small number of iterations, fixed. There does not, therefore, appear to be significant improvements possible over consistently taking

$$S(i) = \mathring{S}(i) = \{1,\dots,m\}, \quad \mathring{V}(i+1) = \emptyset, \quad \text{and} \quad V(i+1) = \{1,\dots,m\}$$

(or the converse dual-first order).

Regarding stochastic algorithms, we start with a few options for the sampling of $S(i)$ in Algorithm 2 with iteration-independent probabilities $\pi_{j,i} \equiv \pi_j$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

EXAMPLE 3.12 (Independent probabilities). If all the blocks $\{1,\dots,m\}$ are chosen independently, we have $\mathbb{P}(\{j,k\} \subset S(i)) = \pi_j\pi_k$ for $j \neq k$, where $\pi_j \in (0,1]$.
EXAMPLE 3.13 (Fixed number of random blocks). If we have a fixed number M of processors, then we may want to choose a subset $S(i) \subset \{1,\dots,m\}$ such that $\#S(i) = M$.

The next example gives a simple way to satisfy ($\mathcal{V}$.a) for Algorithm 1.
EXAMPLE 3.14 (Alternating x-y and y-x steps). Let us randomly alternate between $\mathring{S}(i) = \emptyset$ and $\mathring{V}(i+1) = \emptyset$. That is, with some probability $\mathtt{p}_x$, we choose to take an x-y step that omits lines 9 and 8 in Algorithm 1 and with probability $1-\mathtt{p}_x$, an y-x step that omits the lines 7 and 10. If $\widetilde{\pi}_j = \mathbb{P}[j \in \mathring{S} | \mathring{S} \neq \emptyset]$ and $\widetilde{\nu}_\ell = \mathbb{P}[\ell \in \mathring{V} | \mathring{V} \neq \emptyset]$ denote the probabilities of the rule used to sample $\mathring{S} = \mathring{S}(i)$ and $\mathring{V} = \mathring{V}(i+1)$ when non-empty, then ($\mathcal{V}$) gives

$$\mathring{\pi}_j = \mathtt{p}_x\widetilde{\pi}_j, \qquad \pi_j = \mathtt{p}_x\widetilde{\pi}_j + (1-\mathtt{p}_x)\mathbb{P}[j \in \mathcal{V}^{-1}(\mathring{V}) | \mathring{V} \neq \emptyset],$$
$$\mathring{\nu}_\ell = (1-\mathtt{p}_x)\widetilde{\nu}_\ell, \qquad \nu_\ell = (1-\mathtt{p}_x)\widetilde{\nu}_j + \mathtt{p}_x\mathbb{P}[\ell \in \mathcal{V}(\mathring{S}) | \mathring{S} \neq \emptyset].$$

To compute π_j and ν_ℓ we thus need to know $\mathcal{V}$ and the exact sampling pattern.
REMARK 3.15 Based on Example 3.14, we can derive an algorithm where the only randomness comes from alternating between full x-y and full y-x steps.

4. Rates of convergence. We now need to satisfy Assumption 3.2. This involves choosing update rules for η_{i+1}, $\eta^\perp_{\tau,i+1}$, $\eta^\perp_{\sigma,i+1}$, $\phi_{j,i+1}$, and $\psi_{\ell,i+1}$. At the same time, to obtain good convergence rates, we need to make $d^x_{j,N}(\widetilde{\gamma}_j)$ and $d^y_{\ell,N}$ small in (3.10). We perform these tasks here, including stating two final versions of our algorithms, Algorithm 1 (doubly stochastic) and Algorithm 2 (full dual updates). Specifically, in Section 4.1 we introduce and study a deterministic alternative to the example random update rule for $\phi_{j,i+1}$ in Example 3.8. The analysis of the new rule is easier, and it allows the computation of η_i, which will also be deterministic, without communication in parallel implementations of our algorithms. Afterwards, in Section 4.2 we look at possible choices for the parameters $\eta^\perp_{\tau,i}$ and $\eta^\perp_{\sigma,i}$, *which are only needed in stochastic variants of Algorithm 1*. In Sections 4.3–4.6 we then give various useful choices of η_i and $\psi_{\ell,i}$ that yield concrete convergence rates.

We assume for simplicity that the sampling pattern is independent of the iteration,

$$\mathring{\pi}_{j,i} \equiv \mathring{\pi}_j > 0, \qquad \mathring{\nu}_{\ell,i} \equiv \mathring{\nu}_\ell, \qquad \pi_{j,i} \equiv \pi_j, \qquad \text{and} \qquad \nu_{\ell,i} \equiv \nu_\ell. \tag{4.1}$$

4.1. Deterministic primal test updates. The next lemma gives a deterministic alternative to Example 3.8. We recall that $\gamma_j \geq 0$ is the factor of (strong) convexity of G_j.
LEMMA 4.1 *Suppose that Assumptions 3.2(b) and 3.2(d) and also* (4.1) *hold and that $i \mapsto \eta^\perp_{\tau,i}$ is non-decreasing. Suppose, moreover, that either* (3.8a) *holds or $\sup_{j=1,\dots,m}\rho_j = 0$. Also take $\tau_{j,0}, \phi_{j,0} > 0$, and $\bar{\gamma}_j \geq 0$ such that $\rho_j + \bar{\gamma}_j > 0$, and set*

$$\phi_{j,i+1} := \phi_{j,i} + 2(\bar{\gamma}_j\eta_i + \rho_j) \qquad (j = 1,\dots,m; i \in \mathbb{N}). \tag{4.2}$$

Then for some $c_j > 0$ and all $N \geq 1$ it holds that

$$\phi_{j,N+1} \in \mathcal{R}(\mathcal{O}_{N-1}; (0,\infty)), \tag{4.3a}$$
$$\mathbb{E}[\phi_{j,N}] = \phi_{j,0} + 2\rho_j N + 2\bar{\gamma}_j \sum_{i=0}^{N-1} \mathbb{E}[\eta_i], \tag{4.3b}$$
$$\mathbb{E}[\phi^{-1}_{j,N}] \leq c_j N^{-1} \qquad (N \geq 1). \tag{4.3c}$$

If $\widetilde{\gamma}_j \in [\bar{\gamma}_j, \gamma_j]$, $j = 1,\dots,m$, satisfy

$$2\widetilde{\gamma}_j\bar{\gamma}_j\eta_i \leq (\widetilde{\gamma}_j - \bar{\gamma}_j)\delta\phi_{j,i} \qquad (j \in S(i),\ i \in \mathbb{N}), \tag{4.3d}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

then Assumption 3.2(e) holds, and

$$d^x_{j,N}(\widetilde{\gamma}_j) = 18\rho_j C_x N. \tag{4.3e}$$

Finally, if $\eta_i \geq b_j \min_j \phi^p_{j,i}$ for some $p, b_j > 0$, then for some $\widetilde{c}_j > 0$ it holds that

$$1 \geq \bar{\gamma}_j \widetilde{c}_j N^{p+1} \mathbb{E}[\phi^{-1}_{j,N}] \qquad (N \geq 4). \tag{4.3f}$$

Proof. Since Assumption 3.2(b) guarantees that $\eta_i \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$, we deduce (4.3a) from (4.2). In fact, $\phi_{j,i+1}$ is deterministic as long as η_i is deterministic.

The claim (4.3b) follows immediate from (4.2) and

$$\phi_{j,N} = \phi_{j,N-1} + 2(\bar{\gamma}_j \eta_{N-1} + \rho_j) = \phi_{j,0} + 2\rho_j N + 2\bar{\gamma}_j \sum_{i=0}^{N-1} \eta_i. \tag{4.4}$$

Since $i \mapsto \eta_i$ is non-decreasing, clearly $\phi_{j,N} \geq 2N\widetilde{\rho}_j$ for $\widetilde{\rho}_j := \rho_j + \bar{\gamma}_j \eta_0 > 0$. Then $\phi^{-1}_{j,N} \leq \frac{1}{2\widetilde{\rho}_j N}$. Taking the expectation proves (4.3c).

Clearly (4.3f) holds if $\bar{\gamma}_j = 0$, so assume that $\bar{\gamma}_j > 0$. Using the assumption $\eta_i \geq b_j \min_j \phi^p_{j,i}$ and $\phi_{j,i} \geq 2i\widetilde{\rho}_j$, which we just proved in (4.4), we estimate

$$\phi_{j,N} \geq \phi_{j,0} + b_j(2\widetilde{\rho}_j)^p \sum_{i=1}^{N} i^p \geq \phi_{j,0} + b_j(2\widetilde{\rho}_j)^p \int_2^N x^p \, dx \geq \phi_{j,0} + p^{-1} b_j (2\widetilde{\rho}_j)^p (N^{p+1} - 2).$$

Thus $\phi^{-1}_{j,N} \leq 1/(\bar{\gamma}_j \widetilde{c}_j N^{1+p})$ for some $\widetilde{c}_j > 0$. Taking the expectation proves (4.3f).

It remains to prove (4.3e) and Assumption 3.2(e). Abbreviating $\gamma_{j,i} := \bar{\gamma}_j + \rho_j \eta_i^{-1}$, we write $\phi_{j,i+1} = \phi_{j,i} + 2\gamma_{j,i}\eta_i$. Since $i \mapsto \eta^{\perp}_{\tau,i}$ is non-decreasing, Corollary 3.6 gives

$$\mathbb{E}[\phi_{j,i}\hat{\tau}_{j,i} | \mathcal{O}_{i-1}] = \eta_i + \eta^{\perp}_{\tau,i} - \eta^{\perp}_{i-1,\tau} \geq \eta_i. \tag{4.5}$$

Expanding the defining equation (3.7b) of $h_{j,i+2}(\widetilde{\gamma}_j)$ with the help of (4.5) we estimate

$$\begin{aligned} h_{j,i+2}(\widetilde{\gamma}_j) &= 2\mathbb{E}[\gamma_{j,i}\eta_i - \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i} | \mathcal{O}_{i-1}] + 2\alpha_i^{-1} |\mathbb{E}[\gamma_{j,i}\eta_i - \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i} | \mathcal{O}_i]| \\ &\leq 2(\gamma_{j,i} - \widetilde{\gamma}_j)\eta_i + 2\alpha_i^{-1} |\gamma_{j,i}\eta_i - \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i}| \\ &\leq 2(1 + \alpha_i^{-1})\rho_j + 2(\bar{\gamma}_j - \widetilde{\gamma}_j)\eta_i + 2\alpha_i^{-1} |\bar{\gamma}_j \eta_i - \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i}|. \end{aligned}$$

Since (4.3d) implies $\bar{\gamma}_j \leq \widetilde{\gamma}_j$, if also

$$\alpha_i^{-1} |\bar{\gamma}_j \eta_i - \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i}| \leq (\widetilde{\gamma}_j - \bar{\gamma}_j)\eta_i, \tag{4.6}$$

then

$$\mathbb{E}[\max\{0, h_{j,i+2}(\widetilde{\gamma}_j)\}] \leq 2(1 + \alpha_i^{-1})\rho_j. \tag{4.7}$$

We claim (4.6) to hold for

$$\alpha_i := \begin{cases} \min_j \bar{\gamma}_j / (\widetilde{\gamma}_j - \bar{\gamma}_j), & \bar{\gamma}_j \eta_i > \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i}, \\ \min_j (\widetilde{\gamma}_j \mathring{\pi}_j^{-1} + \bar{\gamma}_j)/(\widetilde{\gamma}_j - \bar{\gamma}_j), & \bar{\gamma}_j \eta_i \leq \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i}. \end{cases} \tag{4.8}$$

The case $\bar{\gamma}_j \eta_i > \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i}$ is clear. Otherwise, to justify the case $\bar{\gamma}_j \eta_i \leq \widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i}$, we observe that (4.6) can in this case be rewritten as $\widetilde{\gamma}_j \phi_{j,i}\hat{\tau}_{j,i} \leq (\alpha_i(\widetilde{\gamma}_j - \bar{\gamma}_j) - \bar{\gamma}_j)\eta_i$. With the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

choice of α_i in (4.8), we see this to hold if $\phi_{j,i}\hat{\tau}_{j,i} \leq \mathring{\pi}_j^{-1}\eta_i$. We consider the cases $j \in \mathring{S}(i)$ and $j \in S(i) \setminus \mathring{S}(i)$ separately. In the case $j \in \mathring{S}(i)$, this inequality is immediate from (3.6a) in Assumption 3.2(d) and Lemma 3.5. If $j \in S(i) \setminus \mathring{S}(i)$, then (3.6a) and Assumption 3.2(b) give

$$\phi_{j,i}\hat{\tau}_{j,i}(\pi_j - \mathring{\pi}_j) \leq \eta_{\tau,i}^{\perp} \leq \min_{j'}(\pi_{j'} - \mathring{\pi}_{j'})\eta_i \leq (\pi_j - \mathring{\pi}_j)\eta_i \leq (\pi_j - \mathring{\pi}_j)\mathring{\pi}_j^{-1}\eta_i.$$

In the last step we have used that $\mathring{\pi}_j \in (0, 1]$ by (4.1). This finishes verifying (4.7).

Next, we expand (3.7a), obtaining

$$\begin{aligned} q_{j,i+2}(\widetilde{\gamma}_j) &= \big(2\mathbb{E}[\gamma_{j,i}\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}|\mathcal{O}_i] + 2\alpha_i|\mathbb{E}[\gamma_{j,i}\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}|\mathcal{O}_i]| - \delta\phi_{j,i}\big)\chi_{S(i)}(j) \\ &= \big(2(\gamma_{j,i}\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}) + 2\alpha_i|\gamma_{j,i}\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}| - \delta\phi_{j,i}\big)\chi_{S(i)}(j) \\ &\leq \big(2(1+\alpha_i)\rho_j + 2(\bar{\gamma}_j\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}) + 2\alpha_i|\bar{\gamma}_j\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}| - \delta\phi_{j,i}\big)\chi_{S(i)}(j). \end{aligned}$$

Since η_i and $\phi_{j,i}\tau_{j,i}$ are increasing, if also

$$2(\bar{\gamma}_j\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}) + 2\alpha_i|\bar{\gamma}_j\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}| \leq \delta\phi_{j,i} \qquad (j \in S(i)), \tag{4.9}$$

then

$$\mathbb{E}[q_{j,i+2}(\widetilde{\gamma}_j)] \leq 2(1+\alpha_i)\rho_j. \tag{4.10}$$

Inserting α_i from (4.8), we see (4.9) to follow from (4.3d). Finally, (4.7) and (4.10) show that (3.8b) holds with $\rho_j = 0$. Thus Assumption 3.2(e) holds. From Proposition 3.3 we find now

$$\delta^x_{j,i+2}(\widetilde{\gamma}_j) = 8(1+\alpha_i)\rho_j C_x + 2(1+\alpha_i^{-1})\rho_j C_x.$$

Clearly α_i defined in (4.8) is bounded above and below, so we obtain (4.3e). ∎

4.2. The parameters $\eta_{\tau,i}^{\perp}$ and $\eta_{\sigma,i}^{\perp}$. We now want to satisfy Assumption 3.2 (c-i) for doubly-stochastic methods. As it turns out, the parameters $\eta_{\tau,i}^{\perp}$ and $\eta_{\sigma,i}^{\perp}$ do not have any effect on convergence rates. Here are a few options.

LEMMA 4.2 *Assume (4.1) and that $i \mapsto \eta_i$ is non-decreasing with $\eta_i \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$. Then Assumption 3.2(b) and (c-i) hold and both $i \mapsto \eta_{\tau,i}^{\perp}$ and $i \mapsto \eta_{\sigma,i}^{\perp}$ are non-decreasing if either:*

(i) (Constant rule) We take $\eta_{\tau,i}^{\perp} \equiv \eta_{\tau}^{\perp}$ and $\eta_{\sigma,i}^{\perp} \equiv \eta_{\sigma}^{\perp}$ for constant $\eta_{\sigma}^{\perp}, \eta_{\tau}^{\perp} > 0$ satisfying

$$\eta_0 \cdot \min_j(\pi_j - \mathring{\pi}_j) \geq \eta_{\tau}^{\perp} \qquad \textit{and} \qquad \eta_0 \cdot \min_{\ell}(\nu_\ell - \mathring{\nu}_\ell) \geq \eta_{\sigma}^{\perp}.$$

(ii) (Proportional rule) For some $\alpha \in (0,1)$ we take $\eta_{\tau,i}^{\perp} := \eta_{\sigma,i}^{\perp} := \alpha\eta_i$ satisfying

$$\min_j\ (\pi_j - \mathring{\pi}_j) \geq \alpha \qquad \textit{and} \qquad \min_{\ell}\ (\nu_\ell - \mathring{\nu}_\ell) \geq \alpha.$$

Proof. Clearly both rules satisfy Assumption 3.2(b) and (c-i). That $i \mapsto \eta_{\tau,i}^{\perp}$ and $i \mapsto \eta_{\sigma,i}^{\perp}$ are non-decreasing and belong to $\mathcal{R}(\mathcal{O}_{i-1}; [0,\infty))$ is obvious. ∎

4.3. Worst-case rules for η_i. To verify Assumption 3.2(a) we take deterministic worst-case bounds $\mathbb{W}_j, \mathbb{W}_{j,\ell} \geq 0$ such that

$$\mathbb{W}_j := \max_{\ell} \mathbb{W}_{\ell,j} \quad \text{and} \quad \mathbb{W}_{\ell,j} \geq \mathring{\pi}_j^{-1}\chi_{\mathring{S}(i)}(j) + \mathring{\nu}_\ell^{-1}\chi_{\mathring{V}(i+1)}(\ell) \qquad (i \in \mathbb{N}). \tag{4.11}$$

Since we assume iteration-independent probabilities (4.1), such bounds exist.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

LEMMA 4.3 *Suppose Assumption 3.2(d) and (4.1) hold. With* $(\kappa_1, \dots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$ *take*

$$\eta_i := \min_{\ell=1,\dots,n} \sqrt{\frac{(1-\delta)\psi_{\ell,i+1}}{\kappa_\ell(\mathbb{W}_{\ell,1}^2\phi_{1,i}^{-1}, \dots, \mathbb{W}_{\ell,m}^2\phi_{m,i}^{-1})}} \qquad (i \geq 0). \tag{4.12}$$

Then Assumption 3.2(a) holds. Moreover, we have $\eta_i \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$ *provided that* $\psi_{i+1} \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$.

Proof. Recalling the expression for $\lambda_{\ell,j,i}$ in (3.5), Assumption 3.2(d) and (4.1) imply $\lambda_{\ell,j,i} \leq \eta_i \mathbb{W}_{\ell,j}$ for $\ell \in \mathcal{V}(j)$. By the monotonicity of κ_ℓ (assumed in Definition 2.2), Assumption 3.2(a) will therefore hold if

$$\psi_{\ell,i+1} \geq \frac{\eta_i^2}{1-\delta}\kappa_\ell(\mathbb{W}_{\ell,1}^2\phi_{1,i}^{-1}, \dots, \mathbb{W}_{\ell,m}^2\phi_{m,i}^{-1}).$$

This is verified by inserting η_i from (4.12). Clearly (4.12) also verifies $\eta_i \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$ when $\psi_{i+1} \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$. ∎

The next lemma provides a choice of $\psi_{i+1} \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$ that also satisfies Assumption 3.2(f). The resulting η_i we express below in (4.13) to collect all rules in one place.

LEMMA 4.4 *Let* $(\kappa_1, \dots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$ *and take* η_i *according to (4.12) with* $\psi_{\ell,i+1} = \eta_i^{2-1/p}\psi_{\ell,0}$ *for some* $\psi_{\ell,0} > 0$ *and* $p \in (0,1]$. *If* $\phi_{j,i} \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$, *then* $\eta_i, \psi_{i+1} \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$. *If, moreover, (4.3c) holds, then* $\mathbb{E}[\eta_i] \geq c_\eta^p i^p$ *and* $\eta_i \geq b_\eta^p \min_j \phi_{j,i}^p$ *for some constants* $c_\eta, b_\eta > 0$ *independent of* p.

Proof. That $\eta_i, \psi_{i+1} \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$ is clear from (4.12) and $\phi_{j,i} \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$. With $\underline{\psi}_0 := \min_{\ell=1,\dots,n} \psi_{\ell,0}$ from (4.12) also

$$\eta_i^{1/p} \geq \frac{(1-\delta)\underline{\psi}_0}{\max_{\ell=1,\dots,n} \kappa_\ell(\mathbb{W}_{\ell,1}^2\phi_{1,i}^{-1}, \dots, \mathbb{W}_{\ell,m}^2\phi_{m,i}^{-1})}.$$

Since $\hat{\mu}_{\ell,j,i} = 0$ for $\ell \notin \mathcal{V}(j)$, using Definition 2.2(ii), we get

$$\eta_i^{1/p} \geq \frac{(1-\delta)\underline{\psi}_0}{\overline{\kappa}\sum_{j=1}^n \max_\ell \mathbb{W}_{\ell,j}^2\phi_{j,i}^{-1}} \geq \frac{1}{\sum_{j=1}^n b_j^{-1}\phi_{j,i}^{-1}}$$

for $b_j := (1-\delta)\underline{\psi}_0/(\overline{\kappa}\mathbb{W}_j^2)$. This shows that $\eta_i \geq \min_j b_j^p \phi_{j,i}^p$. Since $x \mapsto 1/x$ and $x \mapsto x^q$ are convex on $[0,\infty)$ for $q \geq 1$, Jensen's inequality gives

$$\mathbb{E}[\eta_i] \geq \frac{1}{\mathbb{E}\big[(\sum_{j=1}^n b_j^{-1}\phi_{j,i}^{-1})^p\big]} \geq \frac{1}{\big(\sum_{j=1}^n b_j^{-1}\mathbb{E}[\phi_{j,i}^{-1}]\big)^p}.$$

By an application of (4.3c) we obtain $\mathbb{E}[\eta_i] \geq c_\eta^p i^p$ for $c_\eta := 1/\sum_{j=1}^m b_j^{-1}c_j$. ∎

4.4. Mixed rates under partial strong convexity. We are finally ready to state our main result and algorithms. We recall that by Lemma 2.6, (PP) is equivalent to (2.16) under the structural conditions of Assumption 2.1. Dividing the updates of (2.16) into individual block updates and taking the step length rules from Assumption 3.2(d), we obtain the steps of the doubly-stochastic method Algorithm 1. If we perform full dual updates, i.e., force Assumption 3.2 (c-ii) and following Lemma 3.10 taking $\mathring{V}(i+1) = \{1, \dots, n\}$ and $\mathring{S}(i) = S(i)$, we get the simpler steps of Algorithm 2. Regarding the updates of the remaining parameters that are not specified directly in the algorithm skeletons, we start with:

ETNA
Kent State University and
Johann Radon Institute (RICAM)

THEOREM 4.5 *Assume the block-separable structure (GF), writing $\gamma_j \geq 0$ for the factor of (strong) convexity of G_j. Let $\delta \in (0,1)$ and $(\kappa_1, \ldots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$ (see Definition 2.2). In Algorithm 1 or Algorithm 2, take*

(i) $\phi_{j,0} > 0$ *freely and* $\phi_{j,i+1} := \phi_{j,i} + 2(\bar{\gamma}_j \eta_i + \rho_j)$ *for some* $\rho_j \geq 0$ *and* $\bar{\gamma}_j \in [0, \gamma_j]$ *with* $\rho_j + \bar{\gamma}_j > 0$.

(ii) $\psi_{\ell,0} > 0$ *freely and* $\psi_{\ell,i} := \psi_{\ell,0} \eta_i^{2-1/p}$ *for some fixed* $p \in [1/2, 1]$.

(iii) $\eta_{\tau,i}^{\perp}, \eta_{\sigma,i}^{\perp} > 0$ *(in Algorithm 1) following Lemma 4.2, and with* $\mathbb{W}_j$ *given by (4.11)*

$$\eta_i := \min_{\ell=1,\ldots,n} \left(\frac{(1-\delta)\psi_{\ell,0}}{\kappa_\ell(\mathbb{W}_{\ell,1}^2 \phi_{1,i}^{-1}, \ldots, \mathbb{W}_{\ell,m}^2 \phi_{m,i}^{-1})} \right)^p. \tag{4.13}$$

Let $\widehat{u} \in H^{-1}(0)$, *i.e., solve (OC), and suppose the following hold:*

(A) $\sup_{j=1,\ldots,m} \rho_j = 0$ *or* $\sup_{j=1,\ldots,m;\, i \in \mathbb{N}} \|x_j^{i+1} - \widehat{x}_j\|^2 \leq C_x$ *for a constant* $C_x > 0$.

(B) $p = \frac{1}{2}$ *or both* $\sup_{\ell=1,\ldots,n;\, i \in \mathbb{N}} \|y_\ell^{i+1} - \widehat{y}_\ell\|^2 \leq C_y$ *and* $\bar{\gamma}_{j^*} = 0$ *for some* $j^* \in \{1, \ldots, m\}$.

(C) With $\ell^*(j)$ *and* $\underline{\kappa}$ *given by Definition 2.2, for some* $\widetilde{\gamma}_j \in [\bar{\gamma}_j, \gamma_j]$ *for all* $j = 1, \ldots, m$, *we have the initialisation bound*

$$\widetilde{\gamma}_j = \bar{\gamma}_j = 0 \quad \text{or} \quad \frac{2\widetilde{\gamma}_j \bar{\gamma}_j}{\widetilde{\gamma}_j - \bar{\gamma}_j} \left(\frac{1-\delta}{\underline{\kappa} \mathbb{W}_j} \right)^p \leq \delta \psi_{\ell^*(j),0}^{-p} \phi_{j,0}^{1-p}.$$

Then

$$\begin{aligned} &\sum_{j=1}^m \frac{\delta \widetilde{c}_j \bar{\gamma}_j}{2} \mathbb{E}\big[\|x_j^N - \widehat{x}_j\|\big]^2 + g_{p,N} \\ &\qquad \leq \frac{\|u^0 - \widehat{u}\|_{Z_0 M_0}^2 + 18 C_x (\sum_{j=1}^m \rho_j) N + \sum_{\ell=1}^n \psi_{\ell,0} \big(C_* N^{2p-1} + \delta_*\big)}{2N^{p+1}}, \end{aligned} \tag{4.14}$$

when $N \geq 4$ *and with the weighted gap on the ergodic variables,*

$$g_{p,N} := \begin{cases} c_p \mathcal{G}(\widetilde{x}_N, \widetilde{y}_N), & \text{Algorithm 1, } \widetilde{\gamma}_j \leq \gamma_j/2 \text{ for all } j, \\ c_{*,p} \mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N}), & \text{Algorithm 2, } \widetilde{\gamma}_j \leq \gamma_j/2 \text{ for all } j, \\ 0, & \text{otherwise}. \end{cases}$$

The constants $C_*, \delta_* \geq 0$ *are zero if* $p = 1/2$ *while the constants* $c_p, c_{*,p} > 0$.

REMARK 4.6 If $p = 1/2$, (4.14) yields a mixed $O(1/N^{3/2}) + O(1/N^{1/2})$ convergence rate. If $p = 1$, we get a mixed $O(1/N^2) + O(1/N)$ convergence rate.

REMARK 4.7 Theorem 4.5 is valid (with suitable constants) for general primal update rules as long as (4.3) holds and $i \mapsto \phi_{j,i}$ is non-decreasing. This is the case for the deterministic rule of Lemma 4.1. For the random rule of Example 3.8, the rest of the conditions hold, but we have not been able to verify (4.3f). This has the implication that only the gap estimates hold.

Proof. We use Proposition 3.3, so we need to verify Assumptions 2.1 and 3.2. First of all, ($\mathcal{R}$) follows from the updates rules for the testing and step length parameters that only depend on previous realisations of $S(i)$ and $V(i+1)$. The rest of the conditions of Assumption 2.1 are clear from Lemma 2.6, the derivation of Algorithms 1 and 2 from (2.16), and the requisite nesting condition ($\mathcal{V}$) within the algorithms themselves.

Regarding the requirements (a)–(f) of Assumption 3.2, we proceed as follows:

(a) The choice $\psi_{\ell,i+1} := \eta_i^{2-1/p} \psi_{\ell,0}$ in (ii) shows that (4.13) is equivalent to the formula (4.12) for η_i. Thus Lemma 4.3 verifies (a).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 1 Doubly-stochastic primal-dual method.

Require: $K \in \mathcal{L}(X;Y)$, $G \in \mathcal{C}(X)$, and $F^* \in \mathcal{C}(Y)$ with the separable structures (GF).
Require: Rules for $\phi_{j,i}$, $\psi_{\ell,i+1}$, $\eta_{i+1}, \eta^{\perp}_{\tau,i+1}, \eta^{\perp}_{\sigma,i+1} > 0$ (Theorem 4.5, Corollary 4.8, or 4.9).
Require: Sampling patterns for $S(i), \mathring{S}(i), V(i+1)$, and $\mathring{V}(i+1)$ $(i \in \mathbb{N})$ subject to the nesting condition ($\mathcal{V}$) (p. 20) with iteration-independent probabilities (4.1); see Section 3.6.

Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.
Initialise $\tau_{j,-1}, \sigma_{\ell,0} := 0$, $(j = 1, \dots, m; \ell = 1, \dots, m)$.
for all $i \geq 0$ **until** a stopping criterion is satisfied **do**
  Sample $\mathring{S}(i) \subset S(i) \subset \{1, \dots, m\}$ and $\mathring{V}(i+1) \subset V(i+1) \subset \{1, \dots, n\}$.
  For each $j \in \mathring{S}(i)$, compute

$$\tau_{j,i} := \frac{\eta_i - \phi_{j,i-1}\tau_{j,i-1}\chi_{S(i-1)\setminus\mathring{S}(i-1)}(j)}{\phi_{j,i}\mathring{\pi}_{j,i}} \quad \text{and}$$
$$x_j^{i+1} := (I + \tau_{j,i}\partial G_j)^{-1}\left(x_j^i - \tau_{j,i}\textstyle\sum_{\ell\in\mathcal{V}(j)} K_{\ell,j}^* y_\ell^i\right), \quad \text{where} \quad K_{\ell,j} := Q_\ell K P_j.$$

  For each $\ell \in \mathring{V}(i+1)$, compute

$$\sigma_{j,i+1} := \frac{\eta_i - \psi_{j,i}\sigma_{j,i}\chi_{V(i)\setminus\mathring{V}(i)}(j)}{\psi_{j,i+1}\mathring{\nu}_{\ell,i+1}} \quad \text{and}$$
$$y_\ell^{i+1} := (I + \sigma_{\ell,i+1}\partial F_\ell^*)^{-1}\left(y_\ell^i + \sigma_{\ell,i+1}\textstyle\sum_{j\in\mathcal{V}^{-1}(\ell)} K_{\ell,j} x_j^i\right).$$

  For each $j \in \mathring{S}(i)$ and $\ell \in \mathcal{V}(j)$, set

$$\widetilde{w}_{\ell,j}^{i+1} := \theta_{\ell,j,i+1}(x_j^{i+1} - x_j^i) + x_j^{i+1} \qquad \text{with} \quad \theta_{\ell,j,i+1} := \frac{\tau_{j,i}\phi_{j,i}}{\sigma_{\ell,i+1}\psi_{\ell,i+1}}.$$

  For each $\ell \in \mathring{V}(i+1)$ and $j \in \mathcal{V}^{-1}(\ell)$, set

$$\widetilde{v}_{\ell,j}^{i+1} := b_{\ell,j,i+1}(y_\ell^{i+1} - y_\ell^i) + y_\ell^i \qquad \text{with} \quad b_{\ell,j,i+1} := \frac{\sigma_{\ell,i+1}\psi_{\ell,i+1}}{\tau_{j,i}\phi_{j,i}}.$$

  For each $j \in S(i) \setminus \mathring{S}(i)$, compute

$$\tau_{j,i} := \frac{\eta^{\perp}_{\tau,i}}{\phi_{j,i}(\pi_{j,i} - \mathring{\pi}_{j,i})} \quad \text{and}$$
$$x_j^{i+1} := (I + \tau_{j,i}\partial G_j)^{-1}\left(x_j^i - \tau_{j,i}\textstyle\sum_{\ell\in\mathcal{V}(j)} K_{\ell,j}^* \widetilde{v}_{\ell,j}^{i+1}\right).$$

  For each $\ell \in V(i+1) \setminus \mathring{V}(i+1)$ compute

$$\sigma_{j,i+1} := \frac{\eta^{\perp}_{\sigma,i}}{\psi_{j,i+1}(\nu_{\ell,i+1} - \mathring{\nu}_{\ell,i+1})} \quad \text{and}$$
$$y_\ell^{i+1} := (I + \sigma_{\ell,i+1}\partial F_\ell^*)^{-1}\left(y_\ell^i + \sigma_{\ell,i+1}\textstyle\sum_{j\in\mathcal{V}^{-1}(\ell)} K_{\ell,j}\widetilde{w}_{\ell,j}^{i+1}\right).$$

end for

(b) It is clear that $i \mapsto \phi_{j,i}$ and $i \mapsto \psi_{\ell,i}$ are non-decreasing. Therefore (4.12) shows that $i \mapsto \eta_i$ is non-decreasing. Moreover, Lemma 4.4 verifies that $\eta_i \in \mathcal{R}(\mathcal{O}_{i-1}; (0,\infty))$. Algorithm 2 by construction satisfies Assumption 3.2 (c-ii) and has both $\mathring{V}(i+1) = \emptyset$ and $V(i+1) = \{1, \dots, n\}$. It therefore suffices to refer to Lemma 3.10.
Algorithm 1, by its own statement, satisfies (4.1). Therefore, Lemma 4.2 shows Assumption 3.2(b) and (c-i) and that also $i \mapsto \eta^{\perp}_{\tau,i}$ is non-decreasing.
(c) Proved together with (b) above.
(d) These choices are encoded into Algorithm 1. For Algorithm 2 we recall Lemma 3.10.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 2 Block-stochastic primal-dual method, primal randomisation only

Require: $K \in \mathcal{L}(X;Y)$, $G \in \mathcal{C}(X)$, and $F^* \in \mathcal{C}(Y)$ with the separable structures (GF).
Require: Rules for $\phi_{j,i}, \psi_{\ell,i+1}, \eta_{i+1} \in \mathcal{R}(\mathcal{O}_i; (0,\infty))$ (Theorem 4.5, Corollary 4.8, or 4.9).
Require: Iteration-independent (4.1) sampling pattern for the set $S(i)$ ($i \in \mathbb{N}$); see Section 3.6.

1: Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.
2: **for all** $i \geq 0$ **until** a stopping criterion is satisfied **do**
3: Sample $S(i) \subset \{1, \dots, m\}$.
4: For each $j \notin S(i)$, set $x_j^{i+1} := x_j^i$.
5: For each $j \in S(i)$, with $\tau_{j,i} := \eta_i \pi_{j,i}^{-1} \phi_{j,i}^{-1}$, compute

$$x_j^{i+1} := (I + \tau_{j,i} \partial G_j)^{-1} \left(x_j^i - \tau_{j,i} \textstyle\sum_{\ell \in \mathcal{V}(j)} K_{\ell,j}^* y_\ell^i \right), \quad \text{where} \quad K_{\ell,j} := Q_\ell K P_j.$$

6: For each $j \in S(i)$ set

$$\bar{x}_j^{i+1} := \theta_{j,i+1}(x_j^{i+1} - x_j^i) + x_j^{i+1} \qquad \text{with} \quad \theta_{j,i+1} := \frac{\eta_i}{\pi_{j,i}\eta_{i+1}}.$$

7: For each $\ell \in \{1, \dots, n\}$ using $\sigma_{\ell,i+1} := \eta_{i+1}\psi_{\ell,i+1}^{-1}$, compute

$$y_\ell^{i+1} := (I + \sigma_{\ell,i+1} \partial F_\ell^*)^{-1} \left(y_\ell^i + \sigma_{\ell,i+1} \textstyle\sum_{j \in \mathcal{V}^{-1}(\ell)} K_{\ell,j} \bar{x}_j^{i+1} \right).$$

8: **end for**

(e) We use Lemma 4.1. We have already showed Assumption 3.2(b) and (d). Moreover, the algorithms satisfy the iteration-independent probability assumption (4.1). By (A), either $\sup_j \rho_j = 0$ or (3.8a) holds. We still need to satisfy (4.3d). Using Definition 2.2(iii) in (4.13), we estimate

$$\eta_i \leq \left(\frac{(1-\delta)\psi_{\ell^*(j),0}}{\underline{\kappa}\mathbb{W}_j} \phi_{j,i} \right)^p. \tag{4.15}$$

By (C), therefore, either $\widetilde{\gamma}_j = \bar{\gamma}_j = 0$ or $2\widetilde{\gamma}_j\bar{\gamma}_j\eta_i \leq \delta(\widetilde{\gamma}_j - \bar{\gamma}_j)\phi_{j,0}^{1-p}\phi_{j,i}^p$. By (i), $i \mapsto \phi_{j,i}$, so this gives (4.3d). Lemma 4.1 now shows Assumption 3.2(e).

(f) If $p = 1/2$, by Remark 4.6, $\psi_{\ell,i} \equiv \psi_{\ell,0}$. Therefore (3.9b) holds. If $p \neq 1/2$, the same remark and (A) guarantee (3.9a).

With Assumptions 2.1 and 3.2 now verified, Proposition 3.3 provides the estimate

$$\begin{aligned} &\sum_{k=1}^m \frac{\delta}{2\mathbb{E}[\phi_{k,N}^{-1}]} \cdot \mathbb{E}\left[\|x_k^N - \widehat{x}_k\|\right]^2 + \widetilde{g}_N \\ &\qquad \leq \frac{1}{2}\|u^0 - \widehat{u}\|_{Z_0 M_0}^2 + \sum_{j=1}^m \frac{1}{2} d_{j,N}^x(\widetilde{\gamma}_j) + \sum_{\ell=1}^n \frac{1}{2} d_{\ell,N}^y, \end{aligned} \tag{4.16}$$

where $\widetilde{g}_N$, $d_{j,N}^x(\widetilde{\gamma}_j)$, and $d_{\ell,N}^y$ are given in (3.11) To obtain convergence rates, we still need to further analyse this estimate, mainly ζ_N and $\zeta_{*,N}$ within $\widetilde{g}_N$.

We start with ζ_N and $\zeta_{*,N}$. By Lemma 4.2 for Algorithm 1 and directly by Assumption 3.2 (c-ii) for Algorithm 2, $i \mapsto \eta_{\tau,i}^\perp$ is non-decreasing (as is $i \mapsto \eta_{\sigma,i}^\perp$). We recall the coupling variable $\bar{\eta}_i$ from (3.2). Observe that (4.3c) holds as we have verified the conditions of Lemma 4.1 above. By Corollary 3.6 and Lemma 4.4, therefore, in both cases, (3.2a) and

ETNA
Kent State University and
Johann Radon Institute (RICAM)

(3.2b), for some constant $c_\eta > 0$,

$$\bar\eta_i = \mathbb{E}[\eta_i + \eta^\perp_{\tau,i} - \eta^\perp_{\tau,i-1}] \geq \mathbb{E}[\eta_i] \geq c_\eta^p i^p.$$

Thus we estimate ζ_N from (3.3) as

$$\begin{aligned}\zeta_N &= \sum_{i=0}^{N-1} \bar\eta_i \geq \sum_{i=0}^{N-1} \mathbb{E}[\eta_i] \geq c_\eta^p \sum_{i=0}^{N-1} i^p \geq c_\eta^p \int_0^{N-2} x^p\,dx \\ &\geq \frac{c_\eta^p}{p+1}(N-2)^{p+1} \geq \frac{c_\eta^p}{2^{p+1}(p+1)} N^{p+1} =: c_p N^{p+1} \qquad (N \geq 4).\end{aligned} \tag{4.17}$$

Similarly, for some $c_{*,p} > 0$, the quantity $\zeta_{*,N}$ defined in (3.3) satisfies

$$\zeta_{*,N} \geq \sum_{i=1}^{N-1} \mathbb{E}[\eta_i] \geq \frac{c_\eta^p}{p+1}((N-2)^{p+1} - 1) \geq c_{*,p} N^{p+1} \qquad (N \geq 4). \tag{4.18}$$

If $p = 1/2$, (ii) clearly implies $d^y_{\ell,N} = \mathbb{E}[\psi_{\ell,N} - \psi_{\ell,0}] \equiv 0$. Therefore, we can take $C_*, \delta_* = 0$. Otherwise, since $0 \leq 2 - 1/p \leq 1$, the map $t \mapsto t^{2-1/p}$ is concave. Therefore, using (3.11), (ii), and Jensen's inequality, we deduce

$$\begin{aligned}d^N_{y,\ell} &= \sum_{i=0}^{N-1} 9C_y \mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}] = 9C_y\psi_{\ell,0}(\mathbb{E}[\eta_{N+1}^{2-1/p}] - \mathbb{E}[\eta_1^{2-1/p}]) \\ &\leq 9C_y\psi_{\ell,0}\mathbb{E}[\eta_{N+1}]^{2-1/p}.\end{aligned}$$

The condition (B) provides $j^* \in \{1,\dots,m\}$ with $\gamma_{j^*} = 0$, so that a referral to (4.3b) shows $\mathbb{E}[\phi_{j^*,N}] = \phi_{j^*,0} + 2N\rho_{j^*}$. By (4.15) for some $C_*, \delta_* > 0$ then

$$d^N_{y,\ell} \leq 9C_y\psi_{\ell,0}\left(\frac{(1-\delta)\psi_{\ell^*(j^*),0}}{\underline\kappa \mathbb{W}_{j^*}}\mathbb{E}[\phi_{j^*,i}]\right)^{2p-1} \leq \psi_{\ell,0}(C_* N^{2p-1} + \delta_*). \tag{4.19}$$

Finally, Lemma 4.4 shows $\eta_i \geq b_\eta^p \min_j \phi_{j,i}^p$, $(j = 1,\dots,m)$. Thus (4.3f) and (4.3e) in Lemma 4.1 give $1/\mathbb{E}[\phi_{j,N}^{-1}] \geq \bar\gamma_j \widetilde{c}_j N^{p+1}$, for $N \geq 4$, and $d^x_{j,N}(\widetilde\gamma_j) = 18\rho_j C_x N$. Now (4.14) is immediate by applying these estimates and (4.17)–(4.19) to (4.16). ∎

4.5. Unaccelerated algorithm. If $\rho_j = 0$ and $\bar\gamma_j = \widetilde\gamma_j = 0$ for all $j = 1,\dots,m$, then $\phi_{j,i} \equiv \phi_{j,0}$. Consequently Lemma 4.3 gives $\eta_i \equiv \eta_0$. Recalling ζ_N from (3.3), we see that $\zeta_N = N\eta_0$. Likewise $\zeta_{*,N}$ from (3.3) satisfies $\zeta_{*,N} = (N-1)\eta_0$. Clearly also $d^y_{\ell,N} = 0$ and $d^x_{j,N}(\widetilde\gamma_j) = 0$. Inserting this information into (4.16), we immediately obtain:

COROLLARY 4.8 *Assume the block-separable structure* (GF). *Moroever, let* $\delta \in (0,1)$ *and* $(\kappa_1,\dots,\kappa_n) \in \mathcal{K}(K,\mathcal{P},\mathcal{Q})$. *In Algorithms 1 or 2, take*

(i) $\phi_{j,i} \equiv \phi_{j,0}$ *for some fixed* $\phi_{j,0} > 0$.
(ii) $\psi_{\ell,i} \equiv \psi_{\ell,0}$ *for some fixed* $\psi_{\ell,0} > 0$.
(iii) $\eta_i \equiv \eta_0$ *given by* (4.12) *and (in Algorithm 1)* $\eta^\perp_{\tau,i}, \eta^\perp_{\sigma,i} > 0$ *following Lemma 4.2.*

Then

(I) The iterates of Algorithm 1 satisfy $\mathcal{G}(\widetilde x_N, \widetilde y_N) \leq C_0\eta_0^{-1}/(2N)$, *for* $N \geq 1$.
(II) The iterates of Algorithm 2 satisfy $\mathcal{G}(\widetilde x_{*,N}, \widetilde y_{*,N}) \leq C_0/[2\eta_0(N-1)]$, *for* $N \geq 2$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

4.6. Full primal strong convexity. If G is fully strongly convex, we can naturally derive an $O(1/N^2)$ algorithm.

COROLLARY 4.9 *Assume the block-separable structure (GF), assuming each G_j, $j = 1, \dots, m$, strongly convex with the corresponding factor $\gamma_j > 0$. Let $\delta \in (0,1)$ and $(\kappa_1, \dots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$. In Algorithm 1 or Algorithm 2, take*

(i) $\phi_{j,0} > 0$ freely and $\phi_{j,i+1} := \phi_{j,i}(1 + 2\bar{\gamma}_j \tau_{j,i})$ for some fixed $\bar{\gamma}_j \in (0, \gamma_j)$.
(ii) $\psi_{\ell,0} > 0$ freely and $\psi_{\ell,i} := \psi_{\ell,0}$.
(iii) η_i according to (4.12), and (in Algorithm 1) $\eta^{\perp}_{\tau,i}, \eta^{\perp}_{\sigma,i} > 0$ following Lemma 4.2.

Suppose the initialisation bound in Theorem 4.5(C) holds. Then

$$\sum_{j=1}^{m} \frac{\delta \widetilde{c}_j \bar{\gamma}_j}{2} \mathbb{E}\big[\|x_j^N - \widehat{x}_j\|\big]^2 + \widetilde{g}_{1,N} \leq \frac{\|u^0 - \widehat{u}\|^2_{Z_0 M_0}}{2N^2} \qquad (N \geq 4)$$

for

$$\widetilde{g}_{1,N} := \begin{cases} q_1 \mathcal{G}(\widetilde{x}_N, \widetilde{y}_N), & \text{Algorithm 1, } \widetilde{\gamma}_j \leq \gamma_j/2 \text{ for all } j, \\ q_{*,1} \mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N}), & \text{Algorithm 2, } \widetilde{\gamma}_j \leq \gamma_j/2 \text{ for all } j, \\ 0, & \text{otherwise.} \end{cases}$$

The constants $\widetilde{c}_j > 0$ are provided by Lemma 4.1 while $q_1, q_{*,1} > 0$.

Proof. We adapt the argumentation of Theorem 4.5 for the case $p = 1/2$. Indeed, with this choice, our present assumptions satisfy the conditions of that theorem:

(i) With $\rho_j = 0$ this becomes the present one. Since we take $\bar{\gamma}_j > 0$, $\rho_j + \bar{\gamma}_j > 0$ as required.
(ii) This reduces to the present one with $p = 1/2$.
(iii) This becomes the present one since (4.13) with $p = 1/2$ equals (4.12).
(A) This condition trivially holds since $\rho_j = 0$ for all $j = 1, \dots, m$.
(B) This trivially holds when $p = 1/2$.
(C) This we have assumed.

Since $C_*, \delta_* = 0$ when $p = 1/2$, the estimate (4.14) therefore holds with the right-hand side $C_0/(2N^{1+1/2})$. We need to improve this to $C_0/(2N^2)$ by improving the testing variable estimates.

Indeed, the update rule (4.2) now gives

$$\phi_{j,N} \geq \underline{\phi}_0 + \underline{\gamma} \sum_{i=0}^{N-1} \eta_i \geq \underline{\phi}_0 + \underline{\gamma} \sum_{i=0}^{N-1} \eta_i \qquad \text{with} \quad \underline{\phi}_0 := \min_j \phi_{j,0} > 0.$$

Lemma 4.4 shows $\eta_i^2 \geq \underline{b} \min_j \phi_{j,i}$ for some $\underline{b}$. Therefore $\eta_N^2 \geq \underline{b}\underline{\phi}_0 + \underline{b}\underline{\gamma} \sum_{i=0}^{N-1} \eta_i$. Written in another way this says $\eta_N^2 \geq \widetilde{\eta}_N^2$, where

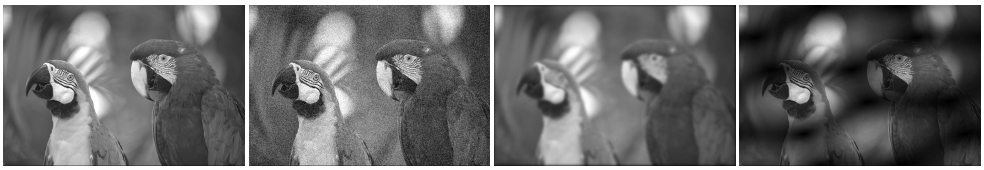
$$\widetilde{\eta}_N^2 = \underline{b}\underline{\phi}_0 + \underline{b}\underline{\gamma} \sum_{i=0}^{N-1} \widetilde{\eta}_i = \widetilde{\eta}_{N-1}^2 + c^2 \underline{\gamma} \widetilde{\eta}_{N-1} = \widetilde{\eta}_{N-1}^2 + \underline{b}\underline{\gamma} \widetilde{\eta}_{N-1}^{-1}.$$

This implies $\eta_i \geq \widetilde{\eta}_i \geq c'_\eta i$ for some $c'_\eta > 0$; cf. the estimates for (2.1) in [6, 37]. Working through the final estimation stage of the proof of Theorem 4.5 with $p = 1/2$, we can now use in (4.17) and (4.18) the estimate $\eta_i \geq c'_\eta i$ that would otherwise correspond to $p = 1$. In our final result, we write the constants c_p and $c_{*,p}$ from the proof as $q_1, q_{*,1} > 0$. ∎

REMARK 4.10 (Linear rates). If both G and F^* are strongly convex, then it is possible to derive linear rates. We refer to [35] for the single-block deterministic case.

REMARK 4.11 (Variance estimates). The variance can be estimated as in [35, Remark 3.4].

ETNA
Kent State University and
Johann Radon Institute (RICAM)



(a) True image. (b) Noisy image. (c) Blurry image. (d) Dimmed image.

FIG. 5.1. *Sample images for denoising, deblurring, and undimming experiments.*

TABLE 5.1
Algorithm variant name construction.

Letter:	1st	2nd	3rd	4th
	Randomisation	ϕ rule	η and ψ rules	κ choice
A-	D: Deterministic P: Primal only B: Primal & Dual	R: Random, Lem. 3.8 D: Determ., Lem. 4.1 C: Constant	B: Bounded: $p = \frac{1}{2}$ I: Increasing: $p = 1$	O: Balanc., Ex. 2.4 M: Max., Ex. 2.3

5. Numerical experience. We now apply several variants of the proposed algorithms to image processing problems. We consider discretisations, as our methods are formulated in Hilbert spaces, but the space of functions of bounded variation—where image processing problems are typically formulated—is only a Banach space. Our specific example problems will be TGV^2 denoising, TV deblurring, and TV undimming.

We present the corrupt and ground-truth images in Figure 5.1, with values in the range $[0, 255]$. We use the images both at the original resolution of $n_1 \times n_2 = 768 \times 512$ and scaled down to 192×128 pixels. To the noisy high-resolution test image in Figure 5.1b, we have added Gaussian noise with a standard deviation of 29.6 (12dB). In the downscaled image, this becomes 6.15 (25.7dB). The image in Figure 5.1c is distorted by Gaussian blur of standard deviation 4. To avoid inverse crimes, we have added Gaussian noise of standard deviation 2.5. The dimmed image in Figure 5.1d is distorted by multiplying the image with a sinusoidal mask γ; see Figure 5.1c. Again, we have added the small amount of noise.

Besides the unaccelerated PDHGM—our examples lack strong convexity for the acceleration of the basic methods—we compare our algorithms to the relaxed PDHGM of [7, 19]. In our precursor work [37], we have also compared these two algorithms to the mixed-rate method of [8] and the adaptive PDHGM of [17]. To keep our tables and figures easily legible, we do not include the algorithms of [37] in our evaluations. It is worth noting that even in the two-block case, the algorithms presented in this paper will not reduce to those of that paper: our rules for $\sigma_{\ell,i}$ are very different from the rules for the single σ_i therein.

We define abbreviations of our algorithm variants in Table 5.1. We do not report the results or apply all variants to all example problems as this would not be informative. We demonstrate the performance of the stochastic variants on TGV^2 denoising only. This merely serves as an example as our problems are not large enough to benefit from being split on a computer cluster, where the benefits of stochastic approaches would be apparent.

To rely on Theorem 4.5 for convergence, we still need to satisfy (3.9a) and (3.8a) or take $\rho_j = 0$. The bound C_y in Assumption 3.2(f) is easily calculated, as in all of our example problems the functional F^* will restrict the dual variable to lie in a ball of known size. The primal variable, on the other hand, is not explicitly bounded. It is, however, possible to prove data-based conservative bounds on the optimal solution; see, e.g., [36, Appendix A]. We can therefore add an artificial bound to the problem to force all iterates to be bounded, replacing G

ETNA
Kent State University and
Johann Radon Institute (RICAM)

by $\widetilde{G}(x) := G(x) + \delta_{B(0,C_x)}(x)$. In practice, to avoid figuring out the exact magnitude of C_x, we update it dynamically. This avoids the constraint from ever becoming active and affecting the algorithm at all. In [36] a "pseudo duality gap" based on this idea was introduced to avoid problems with numerically infinite duality gaps. We will also use them in our reporting: we take the bound C_x as the maximum over all iterations of all tested algorithms and report the duality gap for the problem with $\widetilde{G}$ replacing G. We always report the pseudo-duality gap in decibels $10\log_{10}(\text{gap}^2/\text{gap}_0^2)$ relative to the initial iterate.

In addition to the pseudo-duality gap, we report for each algorithm the distance to a target solution and function value. We report the distance in decibels $10\log_{10}(\|v^i - \widehat{v}\|^2/\|\widehat{v}\|^2)$ and the primal objective value $\text{val}(x) := G(x) + F(Kx)$ relative to the target as $10\log_{10}((\text{val}(x) - \text{val}(\hat{x}))^2/\text{val}(\hat{x})^2)$. The target solution $\widehat{x}$ we compute by taking one million iterations of the PDHGM. We performed our computations with Matlab+C-MEX on a MacBook Pro with 16GB RAM and a 2.8 GHz Intel Core i5 CPU. The initial iterates are $x^0 = 0$ and $y^0 = 0$.

5.1. TGV^2 denoising. In this problem, we write $x = (v, w)$ and $y = (\phi, \psi)$, where v is the image of interest, and take

$$G(x) = \frac{1}{2}\|f - v\|^2, \quad K = \begin{bmatrix} \nabla & -I \\ 0 & \mathcal{E} \end{bmatrix}, \quad \text{and} \quad F^*(y) = \delta_{B(0,\alpha)^{n_1 n_2}}(\phi) + \delta_{B(0,\beta)^{n_1 n_2}}(\psi).$$

Here $\alpha, \beta > 0$ are regularisation parameters, $\mathcal{E}$ is the symmetrised gradient, and the balls are pixelwise Euclidean with the product Π over image pixels. Since there is no further spatial non-uniformity in this problem, it is natural to take as our projections $P_1 x = v$, $P_2 x = w$, $Q_1 y = \phi$, and $Q_2 y = \psi$. It is then not difficult to calculate the optimal κ_ℓ of Example 2.4, so we use only the 'xxxO' variants of the algorithms in Table 5.1.

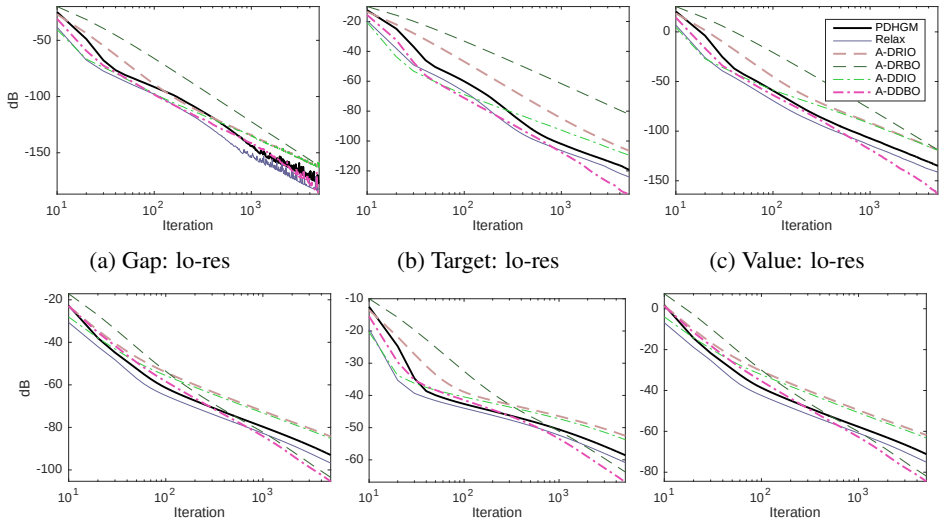
As the regularisation parameters (β, α), we choose $(4.4, 4)$ for the downscaled image. For the original image we scale these parameters by $(0.25^{-2}, 0.25^{-1})$ corresponding to the image downscaling factor [13]. Since G is not strongly convex with respect to w, we have $\widetilde{\gamma}_2 = 0$. For v we take $\widetilde{\gamma}_1 = 1/2$, corresponding to the gap versions of our convergence estimates.

We take $\delta = 0.01$, and parametrise the standard PDHGM with $\sigma_0 = 1.9/\|K\|$ and $\tau_0 \approx 0.52/\|K\|$ solved from $\tau_0\sigma_0 = (1-\delta)\|K\|^2$. These are values that typically work well. For forward-differences discretisation of TGV^2 with cell width $h = 1$, we have $\|K\|^2 \leq 11.4$ [36]. For the 'Relax' method from [7], we use the same σ_0 and τ_0, as well as the value 1.5 for the inertial ρ parameter. For the increasing-ψ 'xxIx' variants of our algorithms, we take $\rho_1 = \rho_2 = 5$, $\tau_{1,0} = \tau_0$, and $\tau_{2,0} = 3\tau_0$. For the bounded-ψ 'xxBx' variants we take $\rho_1 = \rho_2 = 5$, $\tau_{1,0} = \tau_0$, and $\tau_{2,0} = 8\tau_0$. For both methods we also take $\eta_0 = 1/\tau_{0,1}$. *These parametrisations force $\phi_{1,0} = 1/\tau_{1,0}^2$ and keep the initial step length $\tau_{1,0}$ for v consistent with the basic PDHGM.* This justifies our algorithm comparisons using just a single set of parameters. We plot the step length evolution for the A-DDBO variant in Figure 5.3a.

The results for deterministic variants of our algorithm are in Table 5.2 and Figure 5.2. We display the first 5000 iterations in a logarithmic fashion. To reduce computational overheads, we compute the reported quantities only every 10 iterations. To reduce the effects of other processes occasionally occupying the computer, the CPU times reported are the average iteration_time = total_time/total_iterations, excluding time spent initialising the algorithm.

Our first observation is that the variants 'xDxx' based on the deterministic ϕ rule perform better than the "random" rule 'xRxx'. Presently, with no randomisation, the only difference is the value of $\bar{\gamma}$. The value 0.0105 from the initialisation bound in Theorem 4.5(C), for $p = 1/2$, and the value 0.0090, for $p = 1$, appear to give better performance than the maximal value $\widetilde{\gamma}_1 = 0.5$. Generally, the A-DDBO seems to have the best asymptotic performance, with A-DRBO close. A-DDIO has good initial performance, although especially on the higher

ETNA
Kent State University and
Johann Radon Institute (RICAM)



(a) Gap: lo-res (b) Target: lo-res (c) Value: lo-res

(d) Gap: hi-res (e) Target: hi-res (f) Value: hi-res

FIG. 5.2. TGV^2 *denoising, deterministic variants of our algorithms with pixelwise step lengths, 5000 iterations, high (hi-res) and low (lo-res) resolution images.*

TABLE 5.2
TGV^2 *denoising performance: CPU time and number of iterations (at a resolution of 10) to reach a given duality gap, distance to target, or primal objective value.*

low resolution

Method	gap ≤ -60dB iter	time	tgt ≤ -60dB iter	time	val ≤ -60dB iter	time
PDHGM	30	0.21s	100	0.72s	110	0.79s
Relax	20	0.20s	70	0.71s	70	0.71s
A-DRIO	40	0.26s	230	1.55s	180	1.22s
A-DRBO	80	0.54s	890	6.07s	500	3.41s
A-DDIO	20	0.14s	50	0.36s	110	0.80s
A-DDBO	30	0.19s	50	0.32s	90	0.58s

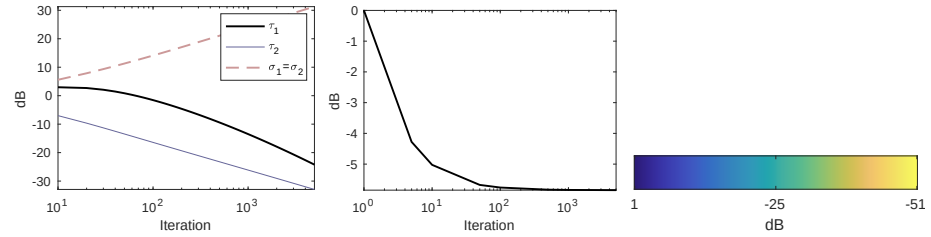
high resolution

gap ≤ -50dB iter	time	tgt ≤ -50dB iter	time	val ≤ -50dB iter	time
50	6.31s	870	111.83s	370	47.49s
40	6.93s	580	102.89s	250	44.25s
70	9.17s	2750	365.52s	1050	139.48s
80	10.56s	860	114.81s	420	56.00s
60	7.37s	2140	267.29s	900	112.34s
60	7.85s	600	79.67s	340	45.09s

resolution image, the PDHGM and 'Relax' perform initially the best. Overall, however, the question of the best performer seems to be a rather fair competition between 'Relax' and A-DDBO.

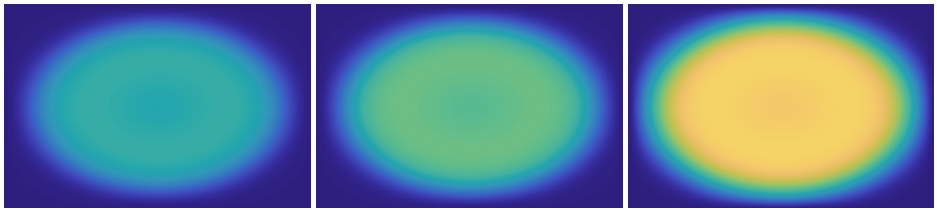
5.2. TGV^2 denoising with stochastic algorithm variants. We also test stochastic variants of our algorithms based on the alternating sampling of Example 3.13 with $M = 1$ and, when appropriate, Example 3.14. We take all probabilities equal to 0.5, that is $\mathrm{p}_x = \widetilde{\pi}_1 =$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



(a) denoising (b) deblurring σ (c) deblurring τ colour coding

(d) deblurring τ, $i = 10$ (e) deblurring τ, $i = 50$ (f) deblurring τ, $i = 100$



(g) deblurring τ, $i = 500$ (h) deblurring τ, $i = 1000$ (i) deblurring τ, $i = 5000$

FIG. 5.3. *Step length evolution (logarithmic from initialisation). A-DDBO* TGV^2 *denoising and A-DDIM* TV *deblurring. The τ plots of the latter are images in the Fourier domain, lighter colour means smaller value of τ* relative to initialisation *(for that specific Fourier component). Note that the images depict logarithm* change*, not absolute values.*

$\widetilde{\pi}_2 = \widetilde{\nu}_1 = \widetilde{\nu}_2 = 0.5$. In the doubly-stochastic 'Bxxx' variants of the algorithms, we take $\eta_{\tau,i}^{\perp} = \eta_{\sigma,i}^{\perp} = 0.9 \cdot 0.5\eta_i$ following the proportional rule Lemma 4.2(ii).

The results are given in Figure 5.4. To conserve space, we have only included a few descriptive algorithm variants. On the x axis, to better describe to the amount of actual work performed by the stochastic methods, the "iteration" count refers to the *expected* number of full primal-dual updates. For all the displayed stochastic variants, with the present choice of probabilities, the expected number of full updates in each iteration is 0.75.

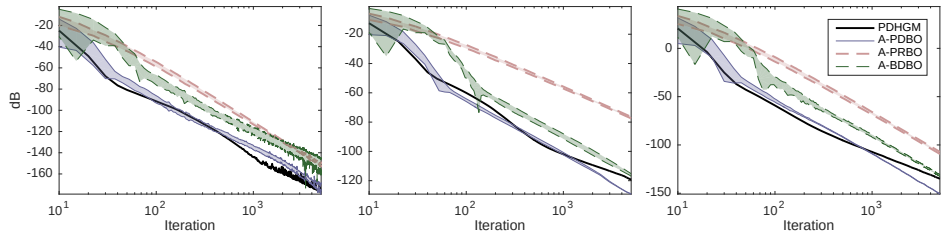
We run each algorithm 50 times and plot for each iteration the 90% confidence interval according to Student's t-distribution. Towards the 5000th iteration, these generally become very narrow, indicating reliability of the random method. Overall, the full-dual-update 'Pxxx' variants perform better than the doubly-stochastic 'Bxxx' variants. In particular, A-PDBO has a performance comparable to or even better than the PDHGM.

5.3. TV deblurring. We want to remove the blur in Figure 5.1c. We do this by taking

$$G(x) = \frac{1}{2}\|f - \mathcal{F}^*(a\mathcal{F}x)\|^2, \quad K = \nabla, \quad \text{and} \quad F^*(y) = \delta_{B(0,\alpha)^{n_1 n_2}}(y),$$

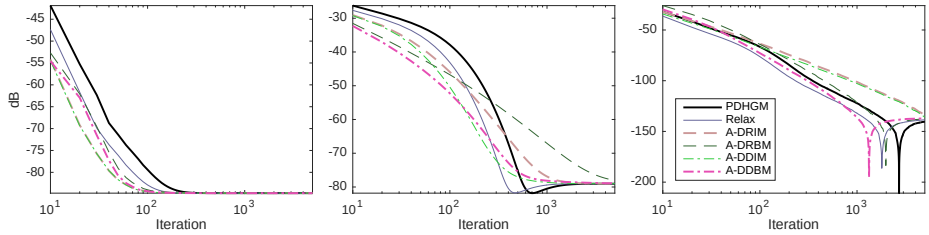
where the balls are again pixelwise Euclidean and with $\mathcal{F}$ the discrete Fourier transform. The factors $a = (a_1, \ldots, a_m)$ model the blurring operation in the Fourier basis.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

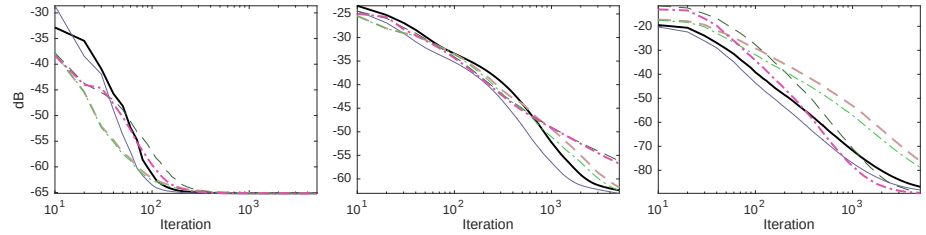


(a) Gap: lo-res (b) Target: lo-res (c) Value: lo-res

FIG. 5.4. TGV^2 *denoising, stochastic variants of our algorithms: 5000 iterations, low resolution images. Iteration number scaled by the fraction of blocks updated on average. For each iteration, 90% confidence interval according to the t-distribution over 50 random runs.*



(a) Gap: lo-res (b) Target: lo-res (c) Value: lo-res



(d) Gap: hi-res (e) Target: hi-res (f) Value: hi-res

FIG. 5.5. TV *deblurring, deterministic variants of our algorithms with pixelwise step lengths, first 5000 iterations, high (hi-res) and low (lo-res) resolution images.*

We take $\alpha = 2.55$ for the high resolution image and scale this to $\alpha = 2.55 * 0.15$ for the low resolution image. We parametrise the PDHGM and 'Relax' algorithms exactly as for TGV^2 denoising above, taking into account the estimate $8 \geq \|K\|^2$ [5]. We take $Q_1 = I$ and P_j as the projection to the jth Fourier component, so $m = n_1 n_2$ and $n = 1$. Thus, *each primal Fourier component has its own step length parameter*. We initialise the latter as $\tau_{j,0} = \tau_0/(\lambda + (1-\lambda)\gamma_j)$ with the componentwise factor of strong convexity $\gamma_j = |a_j|^2$. For the bounded-ψ 'xxBx' algorithm variants we take $\lambda = 0.01$ and for the increasing-ψ 'xxIx' variants $\lambda = 0.1$. We illustrate the step length evolution of the variant A-DDIM in Figure 5.3.

We only experiment with deterministic algorithms as we do not expect small-scale randomisation to be beneficial. We also use the maximal κ 'xxxM' variants, as a more optimal κ would be difficult to compute. The results are presented in Table 5.3 and Figure 5.5. Similarly to A-DDBO in our TGV^2 denoising experiments, A-DDBM performs reliably well, indeed better than the PDHGM or 'Relax'. However, in many cases, A-DRBM and A-DDIM are even faster.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

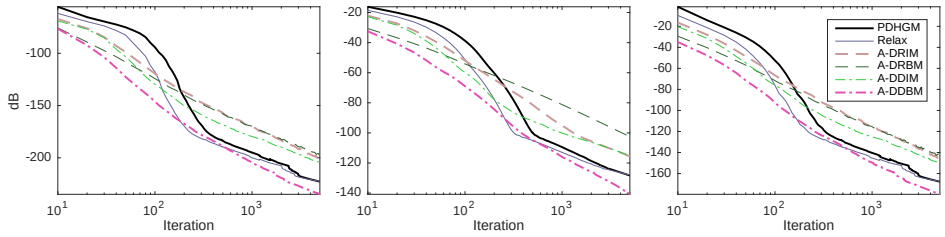
TABLE 5.3
TV *deblurring performance: CPU time and number of iterations (at a resolution of 10) to reach a given duality gap, distance to target, or primal objective value.*

low resolution

	gap ≤ -60dB		tgt ≤ -60dB		val ≤ -60dB	
Method	iter	time	iter	time	iter	time
PDHGM	30	0.18s	330	2.05s	70	0.43s
Relax	20	0.11s	220	1.30s	50	0.29s
A-DRIM	20	0.14s	280	2.08s	80	0.59s
A-DRBM	20	0.14s	490	3.58s	90	0.65s
A-DDIM	20	0.14s	170	1.25s	70	0.51s
A-DDBM	20	0.15s	180	1.37s	60	0.45s

high resolution

gap ≤ -50dB		tgt ≤ -40dB		val ≤ -40dB	
iter	time	iter	time	iter	time
60	5.04s	330	28.12s	110	9.31s
50	4.32s	220	19.30s	90	7.84s
30	3.27s	280	31.41s	320	35.92s
60	6.48s	240	26.27s	220	24.07s
30	3.17s	260	28.35s	230	25.06s
50	5.56s	230	25.98s	150	16.90s



(a) Gap: lo-res (b) Target: lo-res (c) Value: lo-res

(d) Gap: hi-res (e) Target: hi-res (f) Value: hi-res

FIG. 5.6. TV *undimming, deterministic variants of our algorithms with pixelwise step lengths, 5000 iterations, high (hi-res) and low (lo-res) resolution images.*

5.4. TV undimming. We take K and F^* as for TV deblurring but $G(u) := \frac{1}{2}\|f - \gamma \cdot u\|^2$ for the sinusoidal dimming mask $\gamma : \Omega \to \mathbb{R}$. Our experimental setup is also nearly the same as for TV deblurring with the natural difference that the projections P_j are no longer to the Fourier basis but to individual image pixels. The results are presented in Figure 5.6 and Table 5.4. They tell roughly the same story as TV deblurring with A-DDBM performing well and reliably.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 5.4
TV *undimming performance: CPU time and number of iterations (at a resolution of 10) to reach a given duality gap, distance to target, or primal objective value.*

low resolution

	gap ≤ -80dB		tgt ≤ -60dB		val ≤ -60dB	
Method	iter	time	iter	time	iter	time
PDHGM	70	0.18s	200	0.51s	120	0.30s
Relax	50	0.16s	130	0.41s	80	0.25s
A-DRIM	30	0.10s	160	0.57s	80	0.28s
A-DRBM	20	0.05s	170	0.47s	60	0.16s
A-DDIM	30	0.08s	110	0.30s	60	0.16s
A-DDBM	20	0.05s	70	0.18s	40	0.10s

high resolution

gap ≤ -80dB		tgt ≤ -60dB		val ≤ -60dB	
iter	time	iter	time	iter	time
100	3.41s	300	10.31s	210	7.21s
70	3.03s	200	8.73s	140	6.10s
80	3.52s	760	33.82s	640	28.48s
90	3.95s	370	16.39s	380	16.84s
70	3.05s	580	25.57s	430	18.94s
60	2.63s	230	10.22s	200	8.88s

Conclusions. We have derived several accelerated block-proximal primal-dual methods, both stochastic and deterministic. We have concentrated on applying them deterministically, taking advantage of blockwise—indeed pixelwise—factors of strong convexity to obtain improved performance compared to standard methods. In future work, it will be interesting to evaluate the methods on real large scale problems to other state-of-the-art stochastic optimisation methods. Moreover, interesting questions include heuristics and other mechanisms for optimal initialisation of the pixelwise parameters as well as combinations with over-relaxation and inertial schemes such as the extensions of the PDHGM considered in [10, 18, 34, 38].

Acknowledgements. The author would like to thank Peter Richtárik and Olivier Fercoq for several fruitful discussions and for introducing him to stochastic optimisation. Moreover, the support of the EPSRC grant EP/M00483X/1 "Efficient computational tools for inverse imaging problems" is acknowledged during the initial two months of the research.

A data statement for the EPSRC. Implementations of the algorithms described in the paper and relevant boilerplate codes are available on Zenodo at doi:10.5281/zenodo.1042419. The sample photo, also included in the archive, is from the free Kodak image suite at the time of writing at http://r0k.us/graphics/kodak/.

REFERENCES

[1] A. BECK AND M. TEBOULLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.
[2] D. P. BERTSEKAS, *Incremental aggregated proximal and augmented Lagrangian algorithms*, Preprint on arXiv, 2015. https://arxiv.org/abs/1509.09257
[3] P. BIANCHI, W. HACHEM, AND F. IUTZELER, *A stochastic coordinate descent primal-dual algorithm and applications to large-scale composite optimization*, Preprint on arXiv, 2015. https://arxiv.org/abs/1407.0898
[4] J. BOLTE, S. SABACH, AND M. TEBOULLE, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494.
[5] A. CHAMBOLLE, *An algorithm for mean curvature motion*, Interfaces Free Bound., 6 (2004), pp. 195–218.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[6] A. CHAMBOLLE AND T. POCK, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.
[7] ———, *On the ergodic convergence rates of a first-order primal-dual algorithm*, Math. Program., 159 (2016), pp. 253–287.
[8] Y. CHEN, G. LAN, AND Y. OUYANG, *Optimal primal-dual methods for a class of saddle point problems*, SIAM J. Optim., 24 (2014), pp. 1779–1814.
[9] P. L. COMBETTES AND J.-C. PESQUET, *Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration*, Preprint on arXiv, 2016. https://arxiv.org/abs/1602.08021
[10] L. CONDAT, *A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158 (2013), pp. 460–479.
[11] D. CSIBA, Z. QU, AND P. RICHTÁRIK, *Stochastic dual coordinate ascent with adaptive probabilities*, Preprint on arXiv, 2015. https://arxiv.org/abs/1502.08053
[12] I. DAUBECHIES, M. DEFRISE, AND C. DE MOL, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.
[13] J. C. DE LOS REYES, C.-B. SCHÖNLIEB, AND T. VALKONEN, *Bilevel parameter learning for higher-order total variation regularisation models*, J. Math. Imaging Vision, 57 (2017), pp. 1–25.
[14] E. ESSER, X. ZHANG, AND T. F. CHAN, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM J. Imaging Sci., 3 (2010), pp. 1015–1046.
[15] O. FERCOQ AND P. BIANCHI, *A coordinate descent primal-dual algorithm with large step size and possibly non separable functions*, Preprint on arXiv, 2015. https://arxiv.org/abs/1508.04625
[16] O. FERCOQ AND P. RICHTÁRIK, *Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent*, SIAM Rev., 58 (2016), pp. 739–771.
[17] T. GOLDSTEIN, M. LI, AND X. YUAN, *Adaptive primal-dual splitting methods for statistical learning and image processing*, in Advances in Neural Information Processing Systems 28, Vol. 2, C. Cortes, N. D. Lawrence, and D. D. Lee, eds., NIPS, La Jolla, 2015, pp. 2080–2088.
[18] B. HE, Y. YOU, AND X. YUAN, *On the convergence of primal-dual hybrid gradient algorithm*, SIAM J. Imaging Sci., 7 (2014), pp. 2526–2537.
[19] B. HE AND X. YUAN, *Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective*, SIAM J. Imaging Sci., 5 (2012), pp. 119–149.
[20] ———, *Block-wise alternating direction method of multipliers for multiple-block convex programming and beyond*, SMAI J. Comput. Math., 1 (2015), pp. 145–174.
[21] T. MÖLLENHOFF, E. STREKALOVSKIY, M. MOELLER, AND D. CREMERS, *The primal-dual hybrid gradient method for semiconvex splittings*, SIAM J. Imaging Sci., 8 (2015), pp. 827–857.
[22] Y. NESTEROV, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362.
[23] P. OCHS, Y. CHEN, T. BROX, AND T. POCK, *iPiano: inertial proximal algorithm for nonconvex optimization*, SIAM J. Imaging Sci., 7 (2014), pp. 1388–1419.
[24] Z. PENG, T. WU, Y. XU, M. YAN, AND W. YIN, *Coordinate friendly structures, algorithms and applications*, Preprint on arXiv, 2016. https://arxiv.org/abs/1601.00863
[25] Z. PENG, Y. XU, M. YAN, AND W. YIN, *ARock: an algorithmic framework for asynchronous parallel coordinate updates*, SIAM J. Sci. Comput., 38 (2016), pp. A2851–A2879.
[26] J.-C. PESQUET AND A. REPETTI, *A class of randomized primal-dual algorithms for distributed optimization*, J. Nonlinear Convex Anal., 16 (2015), pp. 2453–2490.
[27] T. POCK, D. CREMERS, H. BISCHOF, AND A. CHAMBOLLE, *An algorithm for minimizing the Mumford-Shah functional*, in 12th IEEE Int. Conference on Computer Vision, September 2009, IEEE Conference Proceedings, Los Alamitos, pp. 1133–1140.
[28] Z. QU, P. RICHTÁRIK, AND T. ZHANG, *Randomized dual coordinate ascent with arbitrary sampling*, Preprint on arXiv, 2014. https://arxiv.org/abs/1411.5873
[29] P. RICHTÁRIK AND M. TAKÁČ, *Distributed coordinate descent method for learning with big data*, J. Mach. Learn. Res., 17 (2016), Paper No. 75, (25 pages).
[30] ———, *Parallel coordinate descent methods for big data optimization*, Math. Program., 156 (2016), pp. 433–484.
[31] S. SHALEV-SHWARTZ AND T. ZHANG, *Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization*, Math. Program., 155 (2016), pp. 105–145.
[32] A. N. SHIRYAEV, *Probability*, Graduate Texts in Mathematics, Springer, 1996.
[33] T. SUZUKI, *Stochastic dual coordinate ascent with alternating direction multiplier method*, Preprint on arXiv 2013. https://arxiv.org/abs/1311.0622
[34] T. VALKONEN, *Inertial, corrected, primal–dual proximal splitting*, Preprint on arXiv, 2018. https://arxiv.org/abs/1804.08736
[35] ———, *Testing and non-linear preconditioning of the proximal point method*, Appl. Math. Optim, (2018), in press, https://doi.org/10.1007/s00245-018-9541-6

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[36] T. VALKONEN, K. BREDIES, AND F. KNOLL, *Total generalised variation in diffusion tensor imaging*, SIAM J. Imaging Sci., 6 (2013), pp. 487–525.
[37] T. VALKONEN AND T. POCK, *Acceleration of the PDHGM on partially strongly convex functions*, J. Math. Imaging Vision, 59 (2017), pp. 394–414.
[38] B. C. VŨ, *A splitting algorithm for dual monotone inclusions involving cocoercive operators*, Adv. Comput. Math., 38 (2013), pp. 667–681.
[39] S. J. WRIGHT, *Coordinate descent algorithms*, Math. Program., 151 (2015), pp. 3–34.
[40] A. W. YU, Q. LIN, AND T. YANG, *Doubly stochastic primal-dual coordinate method for empirical risk minimization and bilinear saddle-point problem*, Preprint on arXiv, 2015. https://arxiv.org/abs/1508.03390
[41] Y. ZHANG AND L. XIAO, *Stochastic primal-dual coordinate method for regularized empirical risk minimization*, J. Mach. Learn. Res., 18 (2017), Paper No. 84, (42 pages).
[42] P. ZHAO AND T. ZHANG, *Stochastic optimization with importance sampling*, Preprint on arXiv, 2014. https://arxiv.org/abs/1401.2753
[43] M. ZHU AND T. CHAN, *An efficient primal-dual hybrid gradient algorithm for total variation image restoration*, CAM Report 08-34, Dept. of Math., UCLA, Los Angeles, 2008. ftp://ftp.math.ucla.edu/pub/camreport/cam08-34.pdf