BLOCK-PROXIMAL METHODS WITH SPATIALLY ADAPTED ACCELERATION

Tuomo Valkonen*

V2: 2017-03-16 (V1: 2016-09-16)

Abstract We study and develop (stochastic) primal–dual block-coordinate descent methods based on the method of Chambolle and Pock. Our methods have known convergence rates for the iterates and the ergodic gap: $O(1/N^2)$ if each 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 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.

Get the version from http://tuomov.iki.fi/publications/, citations may be broken in this one due arXiv's inability to support biblatex.

1 INTRODUCTION

We want to efficiently solve optimisation problems of the form

$$\min_x \; G(x) + F(Kx), \tag{1.1}$$

arising from the variational regularisation of 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 functionals on Hilbert spaces X and Y, respectively, and $K \in \mathcal{L}(X; Y)$ to be a bounded linear operator. We are particularly interested in the block-separable case with G and the convex conjugate F^* having the structure

$$G(x) = \sum_{j=1}^m G_j(P_j x), \quad \text{and} \quad F^*(y) = \sum_{\ell=1}^n F^*_\ell(Q_\ell y). \tag{S-GF}$$

Here $P_1, \ldots, P_m$ are projection operators in X with $\sum_{j=1}^m P_j = I$ and $P_j P_i = 0$ if $i \neq j$. Likewise, $Q_1, \ldots, Q_n$ are similarly projection operators in Y.

Several first-order optimisation methods have been developed for (1.1), without the block-separable structure, typically with both G and F convex, and K linear, but recently also accepting

*Department of Mathematical Sciences, University of Liverpool, United Kingdom. tuomo.valkonen@iki.fi

a level of non-convexity and non-linearity [4, 18, 33, 20]. In applications to image processing and data science, one of G or F is typically non-smooth. Effective primal algorithms operating directly on the primal problem (1.1), or its dual, therefore tend to be a form of classical forward–backward splitting, occasionally going by the name of iterative soft-thresholding [11, 1].

In big data optimisation, various forward–backward block-coordinate descent methods have been developed for (1.1) when G block-separable as in (S-GF). At each step of the optimisation method, they only update a subset of the blocks $x_j := P_j x$, randomly in parallel, see e.g. the review [34] and the original articles [19, 26, 15, 27, 25, 37, 28, 10, 9, 22, 2]. Typically F is assumed smooth. Often, each of the functions G_j is assumed strongly convex. Besides parallelism, one advantage of these methods is the exploitation of *local* blockwise factors of smoothness (Lipschitz gradient) of F and K. This can be better than the global factor, and helps convergence.

Unfortunately, primal-only and dual-only stochastic approaches are rarely applicable to image processing and other problems that do not satisfy the separability and smoothness requirements simultaneously, at least not without additional Moreau–Yosida (aka. Huber, aka. Nesterov) regularisation. Generally, even without the splitting into blocks, primal-only or dual-only approaches, as discussed above, can be inefficient on more complicated problems, as the steps of the algorithms become very expensive optimisation problems themselves. This difficulty can often be circumvented through primal-dual approaches. If F is convex, and F^* denotes the conjugate of F, the problem (1.1) can be written

$$\min_x \max_y \; G(x) + \langle Kx, y\rangle - F^*(x). \tag{1.2}$$

If G is also convex, a popular algorithm for (1.2) is the Chambolle–Pock method [6, 24], also classified as the Primal-Dual Hybrid Gradient Method (Modified) or PDHGM in [13]. The method consists of alternating proximal steps on x and y, combined with an over-relaxation step that ensures convergence. It is closely related to the classical ADMM and Douglas–Rachford splitting, as well as the split Bregman method. These connections are discussed in detail in [13].

While early work on block-coordinate descent methods concentrated on primal-only or dual-only algorithms, recently primal-dual algorithms based on the ADMM and the PDHGM have been proposed [30, 36, 14, 3, 21, 23, 35]. Besides [30, 36, 35] that have restrictive smoothness and strong convexity requirements, little is known about the convergence rates of these algorithms.

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, $O(1/N)$ if no convexity is present, and a mixed rate $O(1/N^2) + O(1/N)$ if some of the G_j are strongly convex. These rates apply to an ergodic duality gap, and the faster rates also to the iterates themselves. Our methods will have the additional 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 *both* the primal and dual variable randomly in blocks under a very light compatibility condition. Such "doubly-stochastic" updates, as they are called in [35], have previously been possible only in very limited settings.

Our present work is based on our previous approach in [33] on acceleration of the PDHGM when G is strongly convex only on a subspace. This is the two-block case $m = 2$ and $n = 1$ of (S-GF) with entirely deterministic updates. Besides allowing for (doubly-)stochastic updates and an arbitrary number of both primal *and* dual blocks, in the present work, through a more careful analysis, we derive simplified step length rules.

The more abstract basis of our present work has been split out in [31]. There we study preconditioning of the abstract proximal point method and "testing" by suitable operators as means of obtaining convergence rates. We recall the relevant aspects of this theory in Section 2 along with going through notation and former research on the PDHGM in more detail. In Section 3 we develop the general structure of the promised new methods. We complete the development in Section 4 by deriving step length update rules that yield good convergence rates. We finish with numerical experiments in Section 5. The reader who wishes to skip the detailed derivations may after Section 2 want to go directly to our main result, Theorem 4.5 combined with Algorithms 1 and 2.

2 BACKGROUND AND ABSTRACT RESULTS

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 mean that $T - S$ is positive semidefinite; in particular $T \geq 0$ means that T is positive semidefinite. Also for possibly non-self-adjoint T, we introduce the inner product and norm-like notations

$$\langle x, z\rangle_T := \langle Tx, z\rangle, \quad \text{and} \quad \|x\|_T := \sqrt{\langle x, x\rangle_T}, \tag{2.1}$$

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

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

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

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

2.1 PRIMAL-DUAL ALGORITHMS AS PROXIMAL POINT METHODS

Let us introduce the general variable splitting notation

$$u = (x, y).$$

Following [17, 33], the primal-dual method of Chambolle and Pock [6] (PDHGM) may then in proximal point form be written 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

$$H(u) := \begin{pmatrix} \partial G(x) + K^*y \\ \partial F^*(y) - Kx \end{pmatrix}, \quad \text{and} \quad L_i^0 := \begin{pmatrix} \tau_i^{-1} & -K^* \\ -\widetilde{\omega}_i K & \sigma_{i+1}^{-1} \end{pmatrix}. \tag{2.2}$$

Here $\tau_i, \sigma_{i+1} > 0$ are step length parameters, and $\widetilde{\omega}_i > 0$ an over-relaxation parameter. In the basic version of the algorithm, $\omega_i = 1$, $\tau_i \equiv \tau_0$, and $\sigma_i \equiv \sigma_0$, assuming $\tau_0\sigma_0\|K\|^2 < 1$. 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 that we will return to in Section 2.3.

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

$$\omega_i := 1/\sqrt{1 + 2\widetilde{\gamma}\tau_i}, \quad \tau_{i+1} := \tau_i\omega_i, \quad \text{and} \quad \sigma_{i+1} := \sigma_i/\omega_i. \tag{2.3}$$

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 the ergodic duality gap.

Let then G and F^* have the structure (S-GF). In [33], we extended the PDHGM to partially strongly convex problems: the two-block case $m = 2$ and $n = 1$ with only G_1 strongly convex. This was based on taking in (PP_0) for suitable *invertible* step length operators $T_i \in \mathcal{L}(X; X)$ and $\Sigma_i \in \mathcal{L}(Y; Y)$ the preconditioning operator

$$L_i = \begin{pmatrix} T_i^{-1} & -K^* \\ -\widetilde{\omega}_i K & \Sigma_{i+1}^{-1} \end{pmatrix}. \tag{2.4}$$

In this paper, we want to update any number of blocks stochastically. This will demand the use of *non-invertible* step length operators

$$T_i := \sum_{j \in S(i)} \tau_{j,i} P_j, \quad \text{and} \quad \Sigma_{i+1} := \sum_{\ell \in V(i+1)} \sigma_{\ell,i+1} Q_\ell, \quad (i \geq 0), \tag{S-TΣ}$$

where $\tau_{j,i}, \sigma_{\ell,i+1} \geq 0$ and $S(i) \subset \{1, \ldots, m\}$, $V(i+1) \subset \{1, \ldots, n\}$.

Defining

$$W_{i+1} := \begin{pmatrix} T_i & 0 \\ 0 & \Sigma_{i+1} \end{pmatrix}, \quad \text{and (for now)} \quad M_{i+1} = \begin{pmatrix} I & -T_i K^* \\ -\widetilde{\omega}_i \Sigma_{i+1} K & I \end{pmatrix}, \tag{2.5}$$

the method (PP_0) & (2.4) can also be written

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

This will be the abstract form of our algorithm. To study its convergence, we apply the concept of *testing* that we introduced in [33, 31]. The idea is to analyse the inclusion (PP) by multiplying it with the testing operator

$$Z_{i+1} := \begin{pmatrix} \Phi_i & 0 \\ 0 & \Psi_{i+1} \end{pmatrix} \tag{2.6}$$

for some primal test $\Phi_i \in \mathcal{L}(X; X)$ and dual test $\Psi_{i+1} \in \mathcal{L}(Y; Y)$. To employ the general estimates of [31], we need $Z_{i+1}M_{i+1}$ to be self-adjoint and positive semi-definite. We allow for general $M_{i+1} \in \mathcal{L}(X \times Y; X \times Y)$ instead of the one in (2.5), and assume for some $\Lambda_i \in \mathcal{L}(X; Y)$ that

$$Z_{i+1} M_{i+1} = \begin{pmatrix} \Phi_i & -\Lambda_i^* \\ -\Lambda_i & \Psi_{i+1} \end{pmatrix} \geq 0 \text{ and is self-adjoint}, \quad (i \in \mathbb{N}). \tag{CZ}$$

Expanded, M_{i+1} solved from (CZ), and the proximal maps inverted, (PP) states

$$x^{i+1} = (I + T_i \partial G)^{-1}(x^i + \Phi_i^{-1}\Lambda_i^*(y^{i+1} - y^i) - T_i K^* y^{i+1}), \tag{2.7a}$$

$$y^{i+1} = (I + \Sigma_{i+1}\partial F^*)^{-1}(y^i + \Psi_{i+1}^{-1}\Lambda_i(x^{i+1} - x^i) + \Sigma_{i+1} K x^{i+1}). \tag{2.7b}$$

In an effective algorithm, Λ_i needs to be chosen to avoid cross-dependencies between x^{i+1} and y^{i+1}. An obvious choice would be $\Lambda_i^* = \Phi_i T_i K$. If $V(i+1) \neq \{1, \ldots, n\}$, i.e., for doubly-stochastic algorithms, we will, however, have to make other choices.

Minding the structures (S-GF) and (S-TΣ), in the present work we will take

$$\Phi_i := \sum_{j=1}^{m} \phi_{j,i} P_j, \qquad \Psi_{i+1} := \sum_{\ell=1}^{n} \psi_{\ell,i+1} Q_\ell, \quad \text{and} \tag{S-ΦΨ}$$

$$\Lambda_i := \sum_{j=1}^{m} \sum_{\ell \in \mathcal{V}(j)} \lambda_{\ell,j,i} Q_\ell K P_j, \quad \text{where} \quad \mathcal{V}(j) := \{\ell \in \{1, \ldots, n\} \mid Q_\ell K P_j \neq 0\} \tag{S-Λ}$$

for some $\phi_{j,i}, \psi_{\ell,i+1} > 0$ and $\lambda_{\ell,j,i} \in \mathbb{R}$ over $j = 1, \ldots, m$, $\ell = 1, \ldots, n$, and $i \in \mathbb{N}$. Then Φ_i, Ψ_{i+1}, T_i, and Σ_{i+1} are self-adjoint and positive semi-definite. We also introduce the notation

$$x_j := P_j x, \quad y_\ell := Q_\ell y, \quad \text{and} \quad K_{\ell,j} := Q_\ell K P_j. \tag{2.8}$$

For the moment, we however continue stating background conditions and results for the more convenient abstract structure. In Section 3 we then analyse in detail the block-separable structure, and make specific choices of all the scalar parameters.

2.2 STOCHASTIC SETUP

Just before commencing with the i:th iteration of (PP_0), let us choose T_i and Σ_{i+1} randomly. In practise, we do this through the random choice of $S(i)$ and $V(i+1)$, otherwise based on the information we have gathered before iteration i. This information is modelled by the σ-algebra $\mathcal{O}_{i-1}$, which satisfies $\mathcal{O}_{i-1} \subset \mathcal{O}_i$. To make this formal, let us recall basic measure-theoretic probability from, e.g., [29].

Definition 2.1. 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 also use the next non-standard notation.

Definition 2.2. 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.

To return to our random step length and testing operators, we will assume

$$T_i \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(X;X)), \qquad \Sigma_{i+1} \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(Y;Y)),$$
$$\Phi_i \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(X;X)) \quad \text{and} \qquad \Psi_{i+1} \in \mathcal{R}(\mathcal{O}_i; \mathcal{L}(Y;Y)).$$

We then deduce from (PP) that $x^{i+1} \in \mathcal{R}(\mathcal{O}_i; X)$ and $y^{i+1} \in \mathcal{R}(\mathcal{O}_i; Y)$.

2.3 Ergodic duality gaps and a convergence estimate

We now recall the most central results from our companion paper [31]. To begin to develop duality gaps, we assume for some $\bar\eta_i > 0$ that either

$$
\begin{aligned}
&(\mathrm{C}\mathcal{G}) \qquad \mathbb{E}[T_i^*\Phi_i^*] = \bar\eta_i I, \quad \text{and} \quad \mathbb{E}[\Psi_{i+1}\Sigma_{i+1}] = \bar\eta_i I, \quad (i \ge 1), \quad \text{or}\\
&(\mathrm{C}\mathcal{G}_*) \qquad \mathbb{E}[T_i^*\Phi_i^*] = \bar\eta_i I, \quad \text{and} \quad \mathbb{E}[\Psi_i\Sigma_i] = \bar\eta_i I, \quad (i \ge 1).
\end{aligned}
$$

Correspondingly, with

$$
\zeta_N := \sum_{i=0}^{N-1} \bar\eta_i \quad \text{and} \quad \zeta_{*,N} := \sum_{i=1}^{N-1} \bar\eta_i, \tag{2.9}
$$

we define the ergodic sequences

$$
\widetilde{x}_N := \zeta_N^{-1}\mathbb{E}\left[\sum_{i=0}^{N-1} T_i^*\Phi_i^* x^{i+1}\right], \qquad \widetilde{y}_N := \zeta_N^{-1}\mathbb{E}\left[\sum_{i=0}^{N-1} \Sigma_{i+1}^*\Psi_{i+1}^* y^{i+1}\right], \tag{2.10}
$$

$$
\widetilde{x}_{*,N} := \zeta_{*,N}^{-1}\mathbb{E}\left[\sum_{i=1}^{N-1} T_i^*\Phi_i^* x^{i+1}\right], \qquad \widetilde{y}_{*,N} := \zeta_{*,N}^{-1}\mathbb{E}\left[\sum_{i=1}^{N-1} \Sigma_i^*\Psi_i^* y^{i}\right]. \tag{2.11}
$$

For the accelerated PDHGM, we have $\tau_i\phi_i = \psi\sigma_i$ for a suitable constant ψ and $\phi_i = \tau_i^{-2}$. Therefore ($\mathrm{C}\mathcal{G}_*$) holds while ($\mathrm{C}\mathcal{G}$) does not. In Section 3 we will however see that the latter is the only possibility for doubly-stochastic methods. Introducing

$$
\mathcal{G}(x, y) := \big(G(x) + \langle \widehat{y}, Kx\rangle - F(\widehat{y})\big) - \big(G(\widehat{x}) + \langle y, K\widehat{x}\rangle - F^*(y)\big), \tag{2.12}
$$

the conditions ($\mathrm{C}\mathcal{G}$) and ($\mathrm{C}\mathcal{G}_*$) will then produce two different ergodic duality 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 [31] that forms the basis for our work in the remaining sections. Its statement refers to

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

Theorem 2.1. *Let us be given $K \in \mathcal{L}(X; Y)$, $G \in \mathcal{C}(X)$, and $F^* \in \mathcal{C}(Y)$ with the separable structure (S-GF) on Hilbert spaces X and Y. Denote the factor of (strong) convexity of G_j by $\gamma_j > 0$, and write $\Gamma := \sum_{j=1}^m \gamma_j P_j$. Also let $T_i, \Phi_i \in \mathcal{R}(O_i; \mathcal{L}(X; X))$ and $\Sigma_{i+1}, \Psi_{i+1} \in \mathcal{R}(O_i; \mathcal{L}(Y; Y))$ have the structures (S-TΣ) and (S-ΦΨ). Assuming one of the following conditions and choices of $\widetilde\Gamma$ to hold, let*

$$
\widetilde{g}_N := \begin{cases} 0, & \widetilde\Gamma = \Gamma, \\ \zeta_N \mathcal{G}(\widetilde{x}_N, \widetilde{y}_N), & \widetilde\Gamma = \Gamma/2; (\mathrm{C}\mathcal{G}) \text{ holds}, \\ \zeta_{*,N}\mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N}), & \widetilde\Gamma = \Gamma/2; (\mathrm{C}\mathcal{G}_*) \text{ holds}. \end{cases} \tag{2.13}
$$

Suppose (CZ) holds and that $\Delta_{i+1}(\widetilde\Gamma)$ satisfies

$$
\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}(\Xi_{i+1}(\widetilde\Gamma)+M_{i+1}) - Z_{i+2}M_{i+2}} \ge -\Delta_{i+1}(\widetilde\Gamma). \tag{CΔ}
$$

Then the iterates $u^i = (x^i, y^i)$ of (PP)*, assumed solvable, satisfy*

$$\mathbb{E}\left[\|u^N - \widehat{u}\|^2_{Z_N M_N}\right] + \widetilde{g}_N \le \|u^0 - \widehat{u}\|^2_{Z_1 M_1} + \sum_{i=0}^{N-1} \mathbb{E}[\Delta_{i+1}(\widetilde{\Gamma})]. \tag{2.14}$$

Proof. This is [31, Theorem 4.6] together with [31, Example 4.1]. The latter proves the structure (S-GF), (S-TΣ) & (S-ΦΨ) to satisfy an "ergodic convexity" property that we have avoided introducing here. □

3 BLOCK-PROXIMAL METHODS

The remainder of our work consists of verifying the conditions (CZ), (CΔ) and ($C\mathcal{G}$) or ($C\mathcal{G}_*$) for Theorem 2.1, as well as estimating $\mathbb{E}[\Delta_{i+1}(\widetilde{\Gamma})]$ and $Z_N M_N$. This will be done by refinement of the block-separable step length and testing structure (S-TΣ), (S-ΦΨ) & (S-Λ). Most of this work is done in Sections 3.1 to 3.4, and then combined into almost final algorithms and corresponding convergence results in Section 3.5. We discuss sampling patterns in Section 3.6, and the remaining parameter choices related to convergence rates in Section 4.

For convenience, we introduce

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

The first two denote "effective" step lengths on iteration i. The latter two denote the probability that j will be contained in $S(i)$ and, that ℓ will be contained in $V(i+1)$, given what is known at iteration $i-1$.

3.1 VERIFICATION OF (CZ) AND LOWER BOUNDS FOR $Z_{i+1}M_{i+1}$

The structure (S-ΦΨ) and the choice (2.6) of Z_{i+1} guarantee the self-adjointness of $Z_{i+1}M_{i+1}$ in (CZ) and allow us to solve for M_{i+1}. Since Φ_{i+1} is self-adjoint and positive definite, using (CZ), for any $\delta \in (0,1)$ we moreover deduce

$$Z_{i+1}M_{i+1} = \begin{pmatrix} \Phi_i & -\Lambda_i^* \\ -\Lambda_i & \Psi_{i+1} \end{pmatrix} \ge \begin{pmatrix} \delta\Phi_i & 0 \\ 0 & \Psi_{i+1} - \frac{1}{1-\delta}\Lambda_i\Phi_i^{-1}\Lambda_i^* \end{pmatrix}. \tag{3.1}$$

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

$$(1-\delta)\sum_{\ell=1}^{n} \psi_{\ell,i+1}Q_\ell \ge \sum_{j=1}^{m}\sum_{\ell,k=1}^{n} \lambda_{\ell,j,i}\lambda_{k,j,i}\phi_{j,i}^{-1}Q_\ell K P_j K^* Q_k. \tag{3.2}$$

To go further, we require the functions κ_ℓ introduced next.

Definition 3.1. Let $\mathcal{P} := \{P_1, \dots, P_m\}$, and $Q := \{Q_1, \dots, Q_n\}$. We write $(\kappa_1, \dots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, Q)$ if each $\kappa_\ell : [0,\infty)^m \to [0,\infty)$ is monotone ($\ell = 1, \dots, n$) and for all $\{z_{\ell,j}\} \subset \mathbb{R}$ holds:

(i) (Estimation) The estimate

$$\text{(C-}\kappa\text{.a)}\qquad \sum_{j=1}^{m}\sum_{\ell,k=1}^{n} z_{\ell,j}^{1/2} z_{k,j}^{1/2} Q_\ell K P_j K^* Q_k \le \sum_{\ell=1}^{n} \kappa_\ell(z_{\ell,1},\ldots,z_{\ell,m}) Q_\ell.$$

(ii) (Boundedness) For some $\overline{\kappa} > 0$ the bound

$$\text{(C-}\kappa\text{.b)}\qquad \kappa_\ell(z_1,\ldots,z_m) \le \overline{\kappa}\sum_{j=1}^{m} z_j.$$

(iii) (Non-degeneracy) There exists $\underline{\kappa} > 0$ and $\ell^*(j) \in \{1,\ldots,n\}$ with

$$\text{(C-}\kappa\text{.c)}\qquad \underline{\kappa} z_{j^*} \le \kappa_{\ell^*(j)}(z_1,\ldots,z_m) \quad (j = 1,\ldots,m).$$

Lemma 3.1. *Let* $(\kappa_1,\ldots,\kappa_n) \in \mathcal{K}(K,\mathcal{P},\mathcal{Q})$*, and suppose*

$$\text{(C-}\kappa\psi\text{)}\qquad (1-\delta)\psi_{\ell,i+1} \ge \kappa_\ell(\lambda_{\ell,1,i}^2\phi_{1,i}^{-1},\ldots,\lambda_{\ell,m,i}^2\phi_{m,i}^{-1}) \quad (\ell = 1,\ldots,n).$$

Then (CZ) *holds and*

$$Z_{i+1}M_{i+1} \ge \begin{pmatrix} \delta\Phi_{i+1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.3}$$

Proof. Clearly Φ_{i+1} is self-adjoint and positive definite. The estimate (3.3) follows from (3.2), which follows from (C-κ.a) with $z_{\ell,j} := \lambda_{\ell,j,i}^2\phi_{j,i}^{-1}$. ◻

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

Example 3.1 (Worst-case κ). We may estimate

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

Therefore (C-κ.a) and (C-κ.b) hold with $\overline{\kappa} = \|K\|^2$ for the monotone choice

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

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

Example 3.2 (Balanced κ). Choose a minimal κ_ℓ satisfying (C-κ.a) and the balancing condition

$$\kappa_\ell(z_{\ell,1},\ldots,z_{\ell,m}) = \kappa_k(z_{k,1},\ldots,z_{k,m}) \quad (\ell,k = 1,\ldots,n).$$

This requires problem-specific analysis, but tends to perform well, as we will see in Section 5.

3.2 Verification of (CΔ) and bounds on Δ_{i+1}

For the next lemma, we set

$$A_{i+2} := (\Psi_{i+1}\Sigma_{i+1}K - \Lambda_{i+1}) + (\Lambda_i - KT_i^*\Phi_i^*),$$

and introduce the assumption

$$\mathbb{E}[A_{i+2}|O_i](x^{i+1} - x^i) = 0,\ \mathbb{E}[A_{i+2}^*|O_i](y^{i+1} - y^i) = 0,\ \mathbb{E}[A_{i+2}^*|O_{i-1}] = 0, \tag{CA}$$

which we will seek to enforce in the next Section 3.3. We also recall the coordinate notation x_j and y_ℓ from (2.8).

Lemma 3.2. *Given the structure (S-GF), (S-TΣ), (S-$\Phi\Psi$), and (S-Λ), suppose (CA) holds, and*

$$\mathbb{E}[\psi_{\ell,i+2}|O_i] \geq \mathbb{E}[\psi_{\ell,i+1}|O_i] \quad (\ell = 1,\ldots,n; i \in \mathbb{N}). \tag{C-ψinc}$$

For arbitrary $\alpha_i, \delta > 0$, define

$$\begin{aligned} q_{j,i+2}(\widetilde{\gamma}_j) := \big(&\mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|O_i] \\ &+ \alpha_i|\mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|O_i]| - \delta\phi_{j,i}\big)\chi_{S(i)}(j), \quad \textit{and} \end{aligned} \tag{3.4}$$

$$\begin{aligned} h_{j,i+2}(\widetilde{\gamma}_j) := &\mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|O_{i-1}] \\ &+ \alpha_i^{-1}|\mathbb{E}[\phi_{j,i+1} - \phi_{j,i}(1 + 2\hat{\tau}_{j,i}\widetilde{\gamma}_j)|O_i]|, \end{aligned} \tag{3.5}$$

and assume for some $C_x > 0$ either

$$\|x_j^{i+1} - \widehat{x}_j\|^2 \leq C_x \qquad (j = 1,\ldots,m;\ i \in \mathbb{N}) \quad \textit{or} \tag{C-xbnd.a}$$

$$h_{j,i+2}(\widetilde{\gamma}_j) \leq 0 \quad \textit{and} \quad q_{j,i+2}(\widetilde{\gamma}_j) \leq 0 \qquad (j = 1,\ldots,m;\ i \in \mathbb{N}), \tag{C-xbnd.b}$$

and for some $C_y > 0$ either

$$\|y_\ell^{i+1} - \widehat{y}_\ell\|^2 \leq C_y \qquad (\ell = 1,\ldots,n;\ i \in \mathbb{N}) \quad \textit{or} \tag{C-ybnd.a}$$

$$\mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}|O_i] = 0 \qquad (\ell = 1,\ldots,n;\ i \in \mathbb{N}). \tag{C-ybnd.b}$$

Then (CΔ) is satisfied with $\mathbb{E}[\Delta_{i+1}(\widetilde{\Gamma})] \leq \sum_{j=1}^m \delta^x_{j,i+2}(\widetilde{\gamma}_j) + \sum_{\ell=1}^n \delta^y_{\ell,i+2}$ for

$$\delta^x_{j,i+2}(\widetilde{\gamma}_j) := 4C_x\mathbb{E}[\max\{0, q_{j,i+2}(\widetilde{\gamma}_j)\}] + C_x\mathbb{E}[\max\{0, h_{j,i+2}(\widetilde{\gamma}_j)\}], \quad \textit{and} \tag{3.6}$$

$$\delta^y_{\ell,i+2} := 9C_y\mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}]. \tag{3.7}$$

Proof. The condition (CΔ) holds if we take

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

We then need to estimate $\mathbb{E}[\Delta_{i+1}(\widetilde{\Gamma})]$ from above. Since $u^{i+1} \in \mathcal{R}(O_i; X \times Y)$ and $u^i \in \mathcal{R}(O_{i-1}; X \times Y)$, standard nesting properties of conditional expectations show

$$
\begin{aligned}
\mathbb{E}[\Delta_{i+1}(\widetilde{\Gamma})] = \mathbb{E}\Big[&\|u^{i+1} - u^i\|^2_{\mathbb{E}[D_{i+1}(\widetilde{\Gamma})|O_i]} + \|u^i - \widehat{u}\|^2_{\mathbb{E}[D_{i+1}(\widetilde{\Gamma})|O_{i-1}]} \\
&+ 2\langle u^{i+1} - u^i, u^i - \widehat{u}\rangle_{\mathbb{E}[D_{i+1}(\widetilde{\Gamma})|O_i]}\Big].
\end{aligned} \tag{3.8}
$$

Next we note using (CZ) that

$$
Z_{i+1}(M_{i+1} + \Xi_{i+1}(\widetilde{\Gamma})) = \begin{pmatrix} \Phi_i(I + 2T_i\widetilde{\Gamma}) & 2\Phi_i T_i K^* - \Lambda_i^* \\ -2\Psi_{i+1}\Sigma_{i+1}K - \Lambda_i & \Psi_{i+1} \end{pmatrix}.
$$

In particular, we get

$$
\begin{aligned}
D_{i+1}(\widetilde{\Gamma}) &= \begin{pmatrix} \Phi_{i+1} - \Phi_i(I + 2T_i\widetilde{\Gamma}) & \Lambda_i^* - \Lambda_{i+1}^* - 2\Phi_i T_i K^* \\ 2\Psi_{i+1}\Sigma_{i+1}K + \Lambda_i - \Lambda_{i+1} & \Psi_{i+2} - \Psi_{i+1} \end{pmatrix} \\
&\simeq \begin{pmatrix} \Phi_{i+1} - \Phi_i(I + 2T_i\widetilde{\Gamma}) & A_{i+2}^* \\ A_{i+2} & \Psi_{i+2} - \Psi_{i+1} \end{pmatrix}.
\end{aligned}
$$

Using (CA), we thus expand (3.8) into

$$
\begin{aligned}
\mathbb{E}[\Delta_{i+1}(\widetilde{\Gamma})] = \mathbb{E}\Big[&\|x^{i+1} - x^i\|^2_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|O_i]} + \|y^{i+1} - y^i\|^2_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|O_i]} \\
&+ \|x^i - \widehat{x}\|^2_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|O_{i-1}]} + 2\langle x^{i+1} - x^i, x^i - \widehat{x}\rangle_{\mathbb{E}[\Phi_{i+1} - \Phi_i(I+2T_i\widetilde{\Gamma})|O_i]} \\
&+ \|y^i - \widehat{y}\|^2_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|O_{i-1}]} + 2\langle y^{i+1} - y^i, y^i - \widehat{y}\rangle_{\mathbb{E}[\Psi_{i+2} - \Psi_{i+1}|O_i]}\Big].
\end{aligned} \tag{3.9}
$$

We apply Cauchy's inequality in (3.9) with arbitrary $\alpha_i, \beta_i > 0$. Split into blocks, we obtain $\mathbb{E}[\Delta_{i+1}(\widetilde{\Gamma})] \le \sum_{j=1}^m \delta^x_{j,i+2}(\widetilde{\gamma}_j) + \sum_{\ell=1}^m \delta^y_{\ell,i+2}$ provided for each $j = 1, \ldots, m$ and $i = 1, \ldots, n$, we have the upper bounds

$$
\begin{aligned}
\mathbb{E}\big[q_{j,i+2}(\widetilde{\gamma}_j)\|x_j^{i+1} - x_j^i\|^2 + h_{j,i+2}(\widetilde{\gamma}_j)\|x_j^i - \widehat{x}_j\|^2\big] &\le \delta^x_{j,i+2}(\widetilde{\gamma}_j), \\
\mathbb{E}\big[(1 + \beta_i)\|y_\ell^{i+1} - y_\ell^i\|^2_{\mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}|O_i]} + (1 + \beta_i^{-1})\|y_\ell^i - \widehat{y}_\ell\|^2_{\mathbb{E}[\psi_{\ell,i+2} - \psi_{\ell,i+1}|O_{i-1}]}\big] &\le \delta^y_{\ell,i+2}.
\end{aligned}
$$

These are easy to estimate with (C-xbnd), (C-ybnd), and $\beta_i = 1/2$. □

It is relatively easy to satisfy (C-ψinc) 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 3.3 (Random primal test updates). If (C-xbnd.a) holds, take $\rho_j \ge 0$, otherwise take $\rho_j = 0$ ($j = 1, \ldots, m$). Set

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

Then it is not difficult to show that $\phi_{j,i+1} \in \mathcal{R}(O_i; (0, \infty))$ 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, (R-ϕrnd) is the standard rule (2.3) 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 (R-ϕrnd) is that the coupling parameter η_{i+1} that we introduce in the next section, 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 Lemma 4.1 in Section 4 where we study convergence rates in more detail.

3.3 computability of (PP) and satisfaction of (CA)

As we recall from the discussion after (2.7), we need to choose Λ_i so as to avoid cross-dependencies on x^{i+1} and y^{i+1}. Moreover, we would want $S(i)$ and $V(i+1)$ to correspond exactly to the coordinates x_j^{i+1} and y_j^{i+1} that are indeed updated. We therefore seek to enforce

(C-cons.a) $$x_j^{i+1} = x_j^i, \qquad (j \notin S(i)), \quad \text{and likewise}$$

(C-cons.b) $$y_\ell^{i+1} = y_\ell^i, \qquad (\ell \notin V(i+1)).$$

The next lemma gives a general approach to updating step lengths and sampling blocks such that our demands are met (the condition (C-SV.a) we only use later). To read its statement, we recall $\mathcal{V}$ defined in (S-Λ).

Lemma 3.3. *Assume the structure* (S-GF), (S-TΣ), (S-$\Phi\Psi$), *and* (S-Λ). *The conditions* (CA) *and* (C-cons) *hold if we do the following: For* $i \in \mathbb{N}$, *we take* $\mathring{S}(i)$, $S(i)$, $\mathring{V}(i+1)$, *and* $V(i+1)$ *satisfying*

(C-SV.a) $$\mathcal{V}^{-1}(\mathring{V}(i+1)) \cap \mathcal{V}^{-1}(\mathcal{V}(\mathring{S}(i))) = \emptyset,$$

(C-SV.b) $$S(i) \supset \mathring{S}(i) \cup \mathcal{V}^{-1}(\mathring{V}(i+1)), \quad \textit{and} \quad V(i+1) \supset \mathring{V}(i+1) \cup \mathcal{V}(\mathring{S}(i)),$$

as well as $\eta_i, \eta_{\tau,i}^\perp, \eta_{\sigma,i}^\perp \in \mathcal{R}(O_i; [0,\infty))$ *satisfying*

(C-η) $$i \mapsto \eta_i > 0 \text{ is non-decreasing, and } \begin{cases} \epsilon\eta_i \cdot \min_j(\pi_{j,i} - \mathring{\pi}_{j,i}) \geq \eta_{\tau,i}^\perp, \\ \eta_i \cdot \min_\ell(\nu_{\ell,i} - \mathring{\nu}_{\ell,i}) \geq \eta_{\sigma,i-1}^\perp, \end{cases}$$

for some $\epsilon \in (0,1)$, *independent of* i,

$$\mathring{\pi}_{j,i} := \mathbb{P}[j \in \mathring{S}(i) \mid O_{i-1}], \quad \textit{and} \quad \mathring{\nu}_{\ell,i+1} := \mathbb{P}[\ell \in \mathring{V}(i+1) \mid O_{i-1}].$$

Then, with these assumptions met, we set

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

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

as well as

(R-λ) $$\lambda_{\ell,j,i} := \phi_{j,i}\hat{\tau}_{j,i}\chi_{\mathring{S}(i)}(j) - \psi_{\ell,i+1}\hat{\sigma}_{\ell,i+1}\chi_{\mathring{V}(i+1)}(\ell) \quad (\ell \in \mathcal{V}(j)).$$

Proof. We start by claiming that

$$\mathbb{E}[\lambda_{\ell,j,i+1}|O_i] = \psi_{\ell,i+1}\hat\sigma_{\ell,i+1}(1-\chi_{\mathring{V}(i+1)}(\ell)) - \phi_{j,i}\hat\tau_{j,i}(1-\chi_{\mathring{S}(i)}(j)) \tag{3.10}$$

whenever $\ell \in \mathcal{V}(j)$. Inserting (R-λ), we see (3.10) to be satisfied if

$$\mathbb{E}[\phi_{j,i+1}\hat\tau_{j,i+1}\chi_{\mathring{S}(i+1)}(j)|O_i] = \eta_{i+1} - \phi_{j,i}\hat\tau_{j,i}(1-\chi_{\mathring{S}(i)}(j)) \geq 0, \quad \text{and} \tag{3.11a}$$

$$\mathbb{E}[\psi_{\ell,i+2}\hat\sigma_{\ell,i+2}\chi_{\mathring{V}(i+2)}(\ell)|O_i] = \eta_{i+1} - \psi_{\ell,i+1}\hat\sigma_{\ell,i+1}(1-\chi_{\mathring{V}(i+1)}(\ell)) \geq 0, \tag{3.11b}$$

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

The condition (C-η) guarantees the inequalities in (3.11). To verify the equalities, we observe that the one in (3.11a) can be written

$$\phi_{j,i+1}\mathbb{E}[\tau_{j,i+1}|j \in \mathring{S}(i+1)]\mathring{\pi}_{j,i+1} = \eta_{i+1} - \phi_{j,i}\hat\tau_{j,i}(1-\chi_{\mathring{S}(i)}(j)).$$

If $j \in \mathring{S}(i)$, shifting indices down by one, this is given by the case $j \in \mathring{S}(i)$ of (R-$\tau\sigma$.a). Similarly we cover the case $\ell \in \mathring{V}(i+1)$ of (R-$\tau\sigma$.b). No conditions are set by (3.11) on the remaining cases. However, to cover $j \in S(i) \setminus \mathring{S}(i)$ and $\ell \in V(i+1) \setminus \mathring{V}(i+1)$, we decide to demand

$$\mathbb{E}[\phi_{j,i+1}\hat\tau_{j,i+1}(1-\chi_{\mathring{S}(i+1)}(j))|O_i] = \eta^{\perp}_{\tau,i+1}, \quad \text{and} \tag{3.12a}$$

$$\mathbb{E}[\psi_{\ell,i+2}\hat\sigma_{\ell,i+2}(1-\chi_{\mathring{V}(i+2)}(\ell))|O_i] = \eta^{\perp}_{\sigma,i+1}, \tag{3.12b}$$

These demands are verified by the remaining cases in (R-$\tau\sigma$).

Using (C-SV.b) and (3.10), we observe that $\lambda_{\ell,j,i}$ satisfies

$$\lambda_{\ell,j,i} = 0, \qquad (j \notin S(i) \text{ or } \ell \notin V(i+1)), \quad \text{and} \tag{3.13a}$$

$$\mathbb{E}[\lambda_{\ell,j,i+1}|O_i] = \widetilde{\lambda}_{\ell,j,i+1}, \qquad (j = 1, \dots, m;\ \ell \in \mathcal{V}(j)), \tag{3.13b}$$

where we set $\widetilde{\lambda}_{\ell,j,i+1} := \psi_{\ell,i+1}\hat\sigma_{\ell,i+1} + \lambda_{\ell,j,i} - \phi_{j,i}\hat\tau_{j,i}$. Clearly (3.13a), (S-GF), and (2.7) imply (C-cons). Using (C-cons), (CA) expands as

$$\mathbb{E}[\lambda_{\ell,j,i+1}|O_i] = \widetilde{\lambda}_{\ell,j,i+1} \qquad (j \in S(i),\ \ell \in \mathcal{V}(j)), \tag{3.14a}$$

$$\mathbb{E}[\lambda_{\ell,j,i+1}|O_i] = \widetilde{\lambda}_{\ell,j,i+1} \qquad (\ell \in V(i+1),\ j \in \mathcal{V}^{-1}(\ell)), \quad \text{and} \tag{3.14b}$$

$$\mathbb{E}[\lambda_{\ell,j,i+1}|O_{i-1}] = \mathbb{E}[\widetilde{\lambda}_{\ell,j,i+1}|O_{i-1}], \qquad (j = 1, \dots, m;\ \ell \in \mathcal{V}(j)). \tag{3.14c}$$

Clearly (3.13b) implies (3.14a) and (3.14b). Together with (3.13a), it also implies

$$\mathbb{E}[\mathbb{E}[\lambda_{\ell,j,i+1}|O_i](1-\chi_{V(i+1)}(\ell))(1-\chi_{S(i)}(j))|O_{i-1}] = 0.$$

These allow us to expand

$$\begin{aligned}\mathbb{E}[\lambda_{\ell,j,i+1}|O_{i-1}] &= \mathbb{E}[\mathbb{E}[\lambda_{\ell,j,i+1}|O_i]\chi_{V(i+1)}(\ell)|O_{i-1}] \\ &\quad + \mathbb{E}[\mathbb{E}[\lambda_{\ell,j,i+1}|O_i](1-\chi_{V(i+1)}(\ell))\chi_{S(i)}(j)|O_{i-1}] \\ &\quad + \mathbb{E}[\mathbb{E}[\lambda_{\ell,j,i+1}|O_i](1-\chi_{V(i+1)}(\ell))(1-\chi_{S(i)}(j))|O_{i-1}] \\ &= \mathbb{E}[\widetilde{\lambda}_{\ell,j,i+1}\chi_{V(i+1)}(\ell)|O_{i-1}] + \mathbb{E}[\widetilde{\lambda}_{\ell,j,i+1}(1-\chi_{V(i+1)}(\ell))\chi_{S(i)}(j)|O_{i-1}].\end{aligned} \tag{3.15}$$

On the other hand, (3.13a) implies

$$\mathbb{E}[\widetilde{\lambda}_{\ell,j,i+1}(1-\chi_{V(i+1)}(\ell))(1-\chi_{S(i)}(j))|O_{i-1}] = 0.$$

This and (3.15) show (3.14c). We have therefore verified (CA). □

3.4 SATISFACTION OF THE GAP CONDITIONS ($C\mathcal{G}$) OR ($C\mathcal{G}_*$)

As a corollary of Lemma 3.3, we obtain the following:

Lemma 3.4. *Assume the structure* (S-GF), (S-TΣ), (S-ΦΨ), *and* (S-Λ). *Suppose* (R-$\tau\sigma$) *and* (C-SV) *hold. Then* ($C\mathcal{G}$) *is satisfied if*

$$\mathbb{E}[\eta^\perp_{\tau,i} - \eta^\perp_{\sigma,i}] = \textit{constant}. \tag{C-$\eta^\perp$}$$

Proof. The condition ($C\mathcal{G}$) 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 $\bar\eta_{i+1}$. The updates (R-$\tau\sigma$) (more directly (3.11) and (3.12)) with (C-SV) imply

$$\mathbb{E}[\phi_{j,i+1}\hat\tau_{j,i+1}] = \mathbb{E}[\eta_{i+1} + \eta^\perp_{\tau,i+1} - \eta^\perp_{\tau,i}], \quad \text{and} \tag{3.16a}$$

$$\mathbb{E}[\psi_{\ell,i+2}\hat\sigma_{\ell,i+2}] = \mathbb{E}[\eta_{i+1} + \eta^\perp_{\sigma,i+1} - \eta^\perp_{\sigma,i}]. \tag{3.16b}$$

Thus ($C\mathcal{G}$) follows from (C-$\eta^\perp$). □

Remark 3.5. *If we deterministically take $\mathring{V}(i+1) = \emptyset$, then* (3.12b) *implies $\eta^\perp_{\sigma,i} \equiv 0$. But then* (3.11b) *will be incompatible with* (3.16b). *Therefore $\mathring{V}(i+1)$ has to be random to satisfy* ($C\mathcal{G}$). *The same holds for $\mathring{S}(i)$. Thus algorithms satisfying* ($C\mathcal{G}$) *are necessarily doubly-stochastic, randomly updating both the primal and dual variables, or neither.*

The alternative ($C\mathcal{G}_*$) 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 (3.16a), this holds when $\mathbb{E}[\eta_{i+1} + \eta^\perp_{\tau,i+1} - \eta^\perp_{\tau,i}] = \bar\eta_i = \mathbb{E}[\eta_i + \eta^\perp_{\sigma,i} - \eta^\perp_{\sigma,i-1}]$. It is not clear how to satisfy this simultaneously with (C-η), other than proceeding as in the next lemma.

Lemma 3.6. *Assume the structure* (S-GF), (S-TΣ), (S-ΦΨ), *and* (S-Λ). *Suppose* (R-λ) *holds, and $i \mapsto \eta_i > 0$ is non-decreasing. Take $\eta^\perp_{\tau,i} = 0$, and $\eta^\perp_{\sigma,i} = \eta_{i+1}$ for $i \in \mathbb{N}$. Then* (C-η), (R-$\tau\sigma$), (C-SV), *and* ($C\mathcal{G}_*$) *hold if and only if $\mathring{S}(i) \subset S(i)$, $\mathring{V}(i+1) = \emptyset$, $V(i+1) = \{1,\ldots,n\}$, and*

$$\tau_{j,i} = \eta_i/(\phi_{j,i}\mathring{\pi}_{j,i}) \qquad (j \in S(i)), \tag{R-$\tau\sigma_*$.a}$$

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

Proof. Under our setup, (C-η) holds exactly when $\mathring{\nu}_{\ell,i+1} = 0$, and $\nu_{\ell,i+1} = 1$. This says $\mathring{V}(i+1) = \emptyset$, and $V(i+1) = \{1,\ldots,n\}$. Therefore (C-SV) holds exactly when $S(i) \subset \mathring{S}(i)$. The rule (R-$\tau\sigma_*$) is a specialisation of (R-$\tau\sigma$) to this setup. ($C\mathcal{G}_*$) follows from the discussion above. □

Remark 3.7. *We needed to impose full dual updates to satisfy* ($C\mathcal{G}_*$). *This is akin to most existing primal-dual coordinate descent methods [30, 3, 14]. The algorithms in [23, 21, 35] are more closely related to our method. However only [35] provides convergence rates for very limited single-block sampling schemes under the strong assumption that both G and F^* are strongly convex.*

3.5 BLOCK-PROXIMAL PRIMAL–DUAL ALGORITHMS AND THEIR CONVERGENCE

In the preceding subsections, we have converted the conditions of Theorem 2.1 into more explicit blockwise forms. We collect all of these new conditions in the next proposition. We have not yet specified any of the parameters $\phi_{j,i}, \psi_{\ell,i}, \eta_i, \eta^\perp_{\tau,i}$, or $\eta^\perp_{\sigma,i}$. We will return to these choices in the next section on convergence rates.

Proposition 3.8. *Assume the block-separable structure* (S-GF), (S-TΣ), (S-ΦΨ) *&* (S-Λ). *For each* $j = 1, \ldots, m$*, suppose* G_j *is (strongly) convex with factor* $\gamma_j \geq 0$*, and pick* $\widetilde{\gamma}_j \in [0, \gamma_j]$*. Let* $\delta \in (0, 1)$ *and* $(\kappa_1, \ldots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$*. For each* $i \in \mathbb{N}$*, do the following:*

(i) Sample $\mathring{S}(i) \subset S(i) \subset \{1, \ldots, m\}$ *and* $\mathring{V}(i+1) \subset V(i+1) \subset \{1, \ldots, n\}$ *subject to* (C-SV).

(ii) Define $\tau_{j,i}$ *and* $\sigma_{\ell,i+1}$ *through* (R-$\tau\sigma$)*, and* $\lambda_{\ell,j,i}$ *through* (R-λ).

(iii) Choose $\phi_{j,i}$ *subject to* (C-xbnd)*, and* $\psi_{\ell,i+1}$ *subject to* (C-ybnd), (C-ψinc) *and* (C-$\kappa\psi$).

(iv) Either

  (a) Take η_i*,* $\eta^{\perp}_{\tau,i}$*, and* $\eta^{\perp}_{\sigma,i}$ *satisfying* (C-η) *and* (C-$\eta^{\perp}$)*; or*

  (b) Take $i \mapsto \eta_i > 0$ *non-decreasing,* $\eta^{\perp}_{\tau,i} = 0$*, and* $\eta^{\perp}_{\sigma,i} = \eta_{i+1}$.

Then there exists $C_0 > 0$ *such that the iterates of* (PP) *satisfy* (C-cons) *and*

$$\delta \sum_{k=1}^{m} \frac{1}{\mathbb{E}[\phi^{-1}_{k,N}]} \cdot \mathbb{E}\left[\|x^N_k - \widehat{x}_k\|\right]^2 + \widetilde{g}_N \leq C_0 + \sum_{j=1}^{m} d^x_{j,N}(\widetilde{\gamma}_j) + \sum_{\ell=1}^{n} d^y_{\ell,N}, \tag{3.17}$$

where

$$d^x_{j,N}(\widetilde{\gamma}_j) := \sum_{i=0}^{N-1} \delta^x_{j,i+2}(\widetilde{\gamma}_j), \quad \text{and} \quad d^y_{\ell,N} := \sum_{i=0}^{N-1} \delta^y_{\ell,i+2} = 9C_y \mathbb{E}[\psi_{\ell,N+1} - \psi_{\ell,0}],$$

with $\delta^x_{j,i+2}(\widetilde{\gamma}_j)$ *defined in* (3.6)*, and (see* (2.10)*—*(2.11)*)*

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

Proof. We use Theorem 2.1, so we need to prove (CZ) and (CΔ), as well as the solvability of (PP). For the gap estimates, we also need (C$\mathcal{G}$) or (C$\mathcal{G}_*$). Lemma 3.1 and us assuming (C-$\kappa\psi$) proves (CZ). To prove (CΔ), we proceed as follows: Lemma 3.3 with our assumptions (R-$\tau\sigma$), (R-λ), and (C-SV) proves (CA) as well as (C-cons). The latter applied in (2.7) proves the computability of (PP). Finally, with (CA) verified, Lemma 3.2 and our assumptions (C-xbnd), (C-ybnd), and (C-ψinc) show (CΔ). In case (a), we obtain (C$\mathcal{G}$) from Lemma 3.4, having imposed (R-λ), (C-SV), and (R-$\tau\sigma$). In case (b), Lemma 3.6 similarly yields (C$\mathcal{G}_*$) and (C-η).

It remains to verify (3.17) based on the estimate (2.14) from Theorem 2.1. There we have constrained $\widetilde{\Gamma} = \Gamma$ or $\widetilde{\Gamma} = \Gamma/2$, that is $\widetilde{\gamma}_j \in \{\gamma_j, \gamma_j/2\}$. However, G_j is (strongly) convex with factor γ'_j for any $\gamma'_j \in [0, \gamma_j]$, so we may take $0 \leq \widetilde{\gamma}_j \leq \gamma_j$ with the gap estimates holding when $\widetilde{\gamma}_j \leq \gamma_j/2$. Setting $C_0 := \frac{1}{2}\|u^0 - \widehat{u}\|^2_{Z_0 M_0}$, (2.14) and the estimates of Lemmas 3.1 and 3.2 now yield

$$\delta \mathbb{E}[\|x^N - \widehat{x}\|^2_{\Phi_N}] + \widetilde{g}_N \leq C_0 + \sum_{i=0}^{N-1} (\delta^x_{i+2}(\widetilde{\gamma}_j) + \delta^y_{i+2}).$$

By Hölder's inequality

$$\mathbb{E}[\|x^N - \widehat{x}\|^2_{\Phi_N}] = \sum_{k=1}^{m} \mathbb{E}\left[\phi_{k,N}\|x^N_k - \widehat{x}_k\|^2\right] \geq \sum_{k=1}^{m} \mathbb{E}[\|x^N_k - \widehat{x}_k\|]^2/\mathbb{E}[\phi^{-1}_{k,N}].$$

The estimate (3.17) is now immediate. □

We now write (PP) explicitly in terms of blocks. We already reformulated it in (2.7). We continue from there, first writing $\{\lambda_{\ell,j,i}\}$ from (R-λ) in operator form as

$$\Lambda_i = K\mathring{T}_i^*\Phi_i^* - \Psi_{i+1}\mathring{\Sigma}_{i+1}K, \quad \text{where} \quad \begin{cases} \mathring{T}_i := \sum_{j=1}^m \chi_{\mathring{S}(i)}(j)\hat{\tau}_{j,i}P_j, \text{ and} \\ \mathring{\Psi}_{i+1} := \sum_{j=1}^{\ell} \chi_{\mathring{V}(i+1)}(\ell)\hat{\sigma}_{\ell,i}Q_\ell. \end{cases}$$

Setting $T_i^\perp := T_i - \mathring{T}_i$, and $\Sigma_{i+1}^\perp := \Sigma_{i+1} - \mathring{\Sigma}_{i+1}$, we can thus rewrite (2.7) as

$$v^{i+1} := \Phi_i^{-1}K^*\mathring{\Sigma}^*_{i+1}\Psi^*_{i+1}(y^{i+1} - y^i) + T_i^\perp K^* y^{i+1}, \tag{3.18a}$$

$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - \mathring{T}_i K^* y^i - v^{i+1}), \tag{3.18b}$$

$$z^{i+1} := \Psi_{i+1}^{-1}K\mathring{T}_i^*\Phi_i^*(x^{i+1} - x^i) + \Sigma_{i+1}^\perp K x^{i+1}, \tag{3.18c}$$

$$y^{i+1} := (I + \Sigma_{i+1}\partial F^*)^{-1}(y^i + \mathring{\Sigma}_{i+1}Kx^i + z^{i+1}). \tag{3.18d}$$

Let us set

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

Then thanks to (C-$\mathcal{SV}$), we have $\Sigma_{i+1}^\perp\Theta_{i+1} = \Psi_{i+1}^{-1}K\mathring{T}_i^*\Phi_i^*$. Likewise,

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

satisfies $T_i^\perp B_{i+1}^* = \Phi_i^{-1}K^*\mathring{\Sigma}_{i+1}\Psi_{i+1}$. Now we can rewrite (3.18a) and (3.18c) as

$$v^{i+1} := T_i^\perp[B_{i+1}^*(y^{i+1} - y^i) + K^* y^{i+1}], \quad \text{and} \tag{3.19a}$$

$$z^{i+1} := \Sigma_{i+1}^\perp[\Theta_{i+1}(x^{i+1} - x^i) + Kx^{i+1}]. \tag{3.19b}$$

Observe using (S-GF) how (3.18b) splits with respect to $\mathring{T}_i$ and $T_i^\perp$, while (C-$\mathcal{SV}$.a) guarantees $z^{i+1} = \Sigma_{i+1}^\perp[\Theta_{i+1}(\mathring{x}^{i+1} - x^i) + K\mathring{x}^{i+1}]$. Therefore, (3.18), (3.19) become

$$\mathring{x}^{i+1} := (I + \mathring{T}_i\partial G)^{-1}(x^i - \mathring{T}_i K^* y^i), \tag{3.20a}$$

$$w^{i+1} := \Theta_{i+1}(\mathring{x}^{i+1} - x^i) + \mathring{x}^{i+1}, \tag{3.20b}$$

$$y^{i+1} := (I + \Sigma_{i+1}\partial F^*)^{-1}(y^i + \mathring{\Sigma}_{i+1}Kx^i + \Sigma_{i+1}^\perp w^{i+1}), \tag{3.20c}$$

$$v^{i+1} := B_{i+1}^*(y^{i+1} - y^i) + y^{i+1}, \tag{3.20d}$$

$$x^{i+1} := (I + T_i^\perp\partial G)^{-1}(\mathring{x}^{i+1} - T_i^\perp v^{i+1}). \tag{3.20e}$$

Using the coordinate notation (2.8) and the parameter setup of Lemma 3.3, the iterations (3.20) expand to Algorithm 1. We obtain from Proposition 3.8:

Algorithm 1 Doubly-stochastic primal-dual method

Require: $K \in \mathcal{L}(X;Y)$, $G \in C(X)$, and $F^* \in C(Y)$ with the separable structures (S-GF). Rules for $\phi_{j,i}$, $\psi_{\ell,i+1}$, η_{i+1}, $\eta^{\perp}_{\tau,i+1}$, $\eta^{\perp}_{\sigma,i+1} \in \mathcal{R}(O_i;[0,\infty))$, as well as sampling rules for $\mathring{S}(i)$ and $\mathring{V}(i+1)$, $(j=1,\ldots,m;\ \ell=1,\ldots,n;\ i\in\mathbb{N})$.

Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.

for all $i \geq 0$ **until** a stopping criterion is satisfied **do**

Sample $\mathring{S}(i) \subset S(i) \subset \{1,\ldots,m\}$ and $\mathring{V}(i+1) \subset V(i+1) \subset \{1,\ldots,n\}$ subject to (C-SV).

For each $j \notin S(i)$, set $x^{i+1}_j := x^i_j$.

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

$$\tau_{j,i} := \frac{\eta_i - \phi_{j,i-1}\tau_{j,i-1}\chi_{S(i-1)\setminus\mathring{S}(i-1)}(j)}{\phi_{j,i}\mathring{\pi}_{j,i}}, \quad \text{and}$$
$$x^{i+1}_j := (I+\tau_{j,i}\partial G_j)^{-1}\left(x^i_j - \tau_{j,i}\textstyle\sum_{\ell\in\mathcal{V}(j)} K^*_{\ell,j} y^i_\ell\right).$$

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

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

For each $\ell \notin V(i+1)$, set $y^{i+1}_\ell := y^i_\ell$.

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

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

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

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

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

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

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

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

end for

Corollary 3.9. *Let $\delta \in (0,1)$ and $(\kappa_1,\ldots,\kappa_n) \in \mathcal{K}(K,\mathcal{P},\mathcal{Q})$. Suppose* (C-xbnd), (C-$\psi$inc), (C-ybnd), *and* (C-$\kappa\psi$) *hold along with* (C-$\eta^\perp$), (C-η), *and* (C-SV). *Then Algorithm 1 satisfies* (R-$\tau\sigma$), (R-λ), *and* (3.17) *with $\widetilde{g}_N = \zeta_N \mathcal{G}(\widetilde{x}_N,\widetilde{y}_N)$ when $\widetilde{\gamma}_j \leq \gamma_j/2$ for all j, and $\widetilde{g}_N = 0$ otherwise.*

Using Lemma 3.6, and further enforcing $S(i) = \mathring{S}(i)$, we reduce Algorithm 1 to Algorithm 2.

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 (S-GF). Rules for $\phi_{j,i}, \psi_{\ell,i+1}, \eta_{i+1} \in \mathcal{R}(O_i;(0,\infty))$, as well as a sampling rule for the set $S(i)$, ($j = 1,\dots,m$; $\ell = 1,\dots,n$; $i \in \mathbb{N}$).

Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.

for all $i \geq 0$ **until** a stopping criterion is satisfied **do**

Select random $S(i) \subset \{1,\dots,m\}$.

For each $j \notin S(i)$, set $x_j^{i+1} := x_j^i$.

For each $j \in S(i)$, with $\tau_{j,i} := \eta_i \pi_{j,i}^{-1}\phi_{j,i}^{-1}$, compute

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

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

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

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

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

end for

Its convergence is characterised by:

Corollary 3.10. *Let $\delta \in (0,1)$ and $(\kappa_1,\dots,\kappa_n) \in \mathcal{K}(K,\mathcal{P},\mathcal{Q})$. Suppose (C-xbnd), (C-$\psi$inc), (C-ybnd), and (C-$\kappa\psi$) hold, and $i \mapsto \eta_i > 0$ is non-decreasing. Then Algorithm 2 satisfies (R-$\tau\sigma_*$), (R-λ), and (3.17) with $\widetilde{g}_N = \zeta_{*,N}\mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N})$ when $\widetilde{\gamma}_j \leq \gamma_j/2$ for all j, and $\widetilde{g}_N = 0$ otherwise.*

3.6 Sampling patterns

The only possible fully deterministic sampling patterns allowed by (3.11) and (C-SV) are to consistently take $\mathring{S}(i) = \{1,\dots,m\}$ and $\mathring{V}(i+1) = \emptyset$, or alternatively $\mathring{S}(i) = \emptyset$ and $\mathring{V}(i+1) = \{1,\dots,n\}$. Regarding stochastic algorithms, we start with a few options for sampling $S(i)$ in Algorithm 2 with iteration-independent probabilities $\pi_{j,i} \equiv \pi_j$.

Example 3.4 (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 3.5 (Fixed number of random blocks). If we have a fixed number M of processors, we may want to choose a subset $S(i) \subset \{1,\dots,m\}$ such that $\#S(i) = M$.

The next example gives the simplest way to satisfy (C-SV.a) for Algorithm 1.

Example 3.6 (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 $\mathbb{p}_x$, we choose to take an x-y step that omits lines 11 and 10 in Algorithm 1, and with probability $1 - \mathbb{p}_x$, an y-x step that omits the lines 6 and 9. 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 (C-*SV*) gives

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

To compute π_j and ν_ℓ we thus need to know $\mathcal{V}$ and the exact sampling pattern.

Remark 3.11. *Based on Example 3.6, 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 the conditions of Corollaries 3.9 and 3.10. 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 $d^y_{\ell,N} = \mathbb{E}[\psi_{\ell,N+1} - \psi_{\ell,0}]$ small in (3.17). We do these tasks here. In Section 4.1 we introduce and study a deterministic alternative to the exemplary update rule for $\phi_{j,i+1}$ in Example 3.3. 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 to 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 iteration,

(R-$\pi\nu$) $$\mathring{\pi}_{j,i} \equiv \mathring{\pi}_j > 0, \quad \text{and} \quad \mathring{\nu}_{\ell,i} \equiv \mathring{\nu}_\ell.$$

Then (C-*SV*) shows that also $\pi_{j,i} \equiv \pi_j > 0$ and $\nu_{\ell,i} \equiv \nu_\ell > 0$.

4.1 DETERMINISTIC PRIMAL TEST UPDATES

The next lemma gives a deterministic alternative to Example 3.3. We recall that $\gamma_j \geq 0$ is the factor of strong convexity of G_j.

Lemma 4.1. *Suppose* (R-$\tau\sigma$), (C-η), *and* (R-$\pi\nu$) *hold, and that $i \mapsto \eta^\perp_{\tau,i}$ is non-decreasing. If* (C-xbnd.a) *holds, take $\rho_j \geq 0$, otherwise take $\rho_j = 0$ ($j = 1, \ldots, m$). Also take $\bar{\gamma}_j \geq 0$ such that $\rho_j + \bar{\gamma}_j > 0$, and set*

(R-ϕdet) $$\phi_{j,i+1} := \phi_{j,i} + 2(\bar{\gamma}_j \eta_i + \rho_j), \quad (j = 1, \ldots, m; i \in \mathbb{N}).$$

Then for some $c_j > 0$ holds

$$\phi_{j,N} \in \mathcal{R}(O_{N-1}; (0,\infty)), \tag{4.1a}$$

$$\mathbb{E}[\phi_{j,N}] = \phi_{j,0} + 2\rho_j N + 2\bar{\gamma}_j \sum_{i=0}^{N-1} \mathbb{E}[\eta_i], \quad \textit{and} \tag{4.1b}$$

$$\mathbb{E}[\phi_{j,N}^{-1}] \leq c_j N^{-1}, \quad (N \geq 1). \tag{4.1c}$$

If $\bar{\gamma}_j, \widetilde{\gamma}_j \geq 0$ satisfy

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

then the primal test bound (C-xbnd) holds, and

$$d^x_{j,N}(\widetilde{\gamma}_j) = 18\rho_j C_x N. \tag{4.1d}$$

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

$$\frac{1}{\mathbb{E}[\phi_{j,N}^{-1}]} \geq \bar{\gamma}_j \widetilde{c}_j N^{p+1}, \quad (N \geq 4). \tag{4.1e}$$

Proof. Since $\eta_i \in \mathcal{R}(O_{i-1}; (0,\infty))$, we deduce (4.1a). In fact, $\phi_{j,i+1}$ is deterministic as long as η_i is deterministic. From (R-ϕdet) we compute

$$\phi_{j,N} = \phi_{j,N-1} + 2(\bar{\gamma}_j \eta_{N-1} + \rho_j) = \phi_{j,0} + 2\rho_j N + 2\bar{\gamma}_j \sum_{i=0}^{N-1} \eta_i. \tag{4.2}$$

Since $i \mapsto \eta_i$ is non-decreasing by (C-η), clearly $\phi_{j,N} \geq 2N\widetilde{\rho}_j$ for $\widetilde{\rho}_j := \rho_j + \bar{\gamma}_j \eta_0 > 0$. Then $\phi_{j,N}^{-1} \leq \frac{1}{2\widetilde{\rho}_j N}$. Taking the expectation proves (4.1c), while (4.1b) is immediate from (4.2). Clearly (4.1e) holds if $\bar{\gamma}_j = 0$, so assume $\bar{\gamma}_j > 0$. Using the assumption $\eta_i \geq b_j \min_j \phi^p_{j,i}$ and $\phi_{j,i} \geq 2i\widetilde{\rho}_j$ that we just proved, we get $\eta_i \geq b'_j (i+1)^p$ for some $b'_j > 0$. With this and (4.2) we calculate

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

Thus $\phi_{j,N}^{-1} \leq 1/(\bar{\gamma}_j \widetilde{c}_j N^{1+p})$ for some $\widetilde{c}_j > 0$. Taking the expectation proves (4.1e).

It remains to prove (4.1d) and (C-xbnd). 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, (R-$\tau\sigma$) gives

$$\mathbb{E}[\phi_{j,i}\hat{\tau}_{j,i} | O_{i-1}] = \eta_i + \eta^{\perp}_{\tau,i} - \eta^{\perp}_{i-1,\tau} \geq \eta_i. \tag{4.3}$$

Expanding of (3.5), with the help of (4.3) we estimate

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

Since (C-ϕdet) implies $\bar{\gamma}_j \le \widetilde{\gamma}_j$, it follows

$$\mathbb{E}[\max\{0, h_{j,i+2}(\widetilde{\gamma}_j)\}] \le 2(1+\alpha_i^{-1})\rho_j, \tag{4.4}$$

provided

$$\alpha_i^{-1}|\bar{\gamma}_j\eta_i - \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}| \le (\widetilde{\gamma}_j - \bar{\gamma}_j)\eta_i. \tag{4.5}$$

We claim that this holds for

$$\alpha_i := \begin{cases} \min_j \bar{\gamma}_j/(\widetilde{\gamma}_j - \bar{\gamma}_j), & \bar{\gamma}_j\eta_i > \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}, \\ \max_j(\widetilde{\gamma}_j\mathring{\pi}_j^{-1} + \bar{\gamma}_j)/(\widetilde{\gamma}_j - \bar{\gamma}_j), & \bar{\gamma}_j\eta_i \le \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}. \end{cases} \tag{4.6}$$

The case $\bar{\gamma}_j\eta_i > \widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i}$ is clear. Otherwise, we see that (4.5) holds if

$$\widetilde{\gamma}_j\phi_{j,i}\hat{\tau}_{j,i} \le (\alpha_i(\widetilde{\gamma}_j - \bar{\gamma}_j) - \bar{\gamma}_j)\eta_i. \tag{4.7}$$

Using (4.6), this reduces to the condition $\phi_{j,i}\hat{\tau}_{j,i} \le \mathring{\pi}_j^{-1}\eta_i$. To verify this, we have to consider the cases $j \in \mathring{S}(i)$ and $j \in S(i) \setminus \mathring{S}(i)$ separately. From (R-$\tau\sigma$) we have

$$\phi_{j,i}\hat{\tau}_{j,i}\mathring{\pi}_j\chi_{\mathring{S}(i)}(j) \le \eta_i, \quad \text{and} \quad \phi_{j,i}\hat{\tau}_{j,i}(\pi_j - \mathring{\pi}_j)(1 - \chi_{\mathring{S}(i)}(j)) \le \eta_{\tau,i}^{\perp}.$$

Using (C-η) in the latter estimate, we verify (4.7), and consequently (4.4).

Next, we expand (3.4), obtaining

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

Since η_i and $\phi_{j,i}\tau_{j,i}$ are increasing, we have

$$\mathbb{E}[q_{j,i+2}(\widetilde{\gamma}_j)] \le 2(1+\alpha_i)\rho_j, \tag{4.8}$$

provided

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

Inserting α_i from (4.6), we see that (4.9) follows from (C-ϕdet). Finally, we see from (4.4) and (4.8) that (C-xbnd.b) holds if we take $\rho_j = 0$. Therefore (C-xbnd) always holds. From Lemma 3.2 now

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

Clearly α_i defined in (4.6) is bounded above and below, so we obtain (4.1d). □

4.2 THE PARAMETERS $\eta^\perp_{\tau,i}$ AND $\eta^\perp_{\sigma,i}$

As it turns out, the parameters $\eta^\perp_{\tau,i}$ and $\eta^\perp_{\sigma,i}$, do not have any effect on converge rates, as long as they satisfy (C-$\eta^\perp$) and (C-η). Here we look at a few options.

Lemma 4.2. *Suppose* (R-$\pi\nu$) *holds. The conditions* (C-$\eta^\perp$) *and* (C-η) *hold, and both* $i \mapsto \eta^\perp_{\tau,i}$ *and* $i \mapsto \eta^\perp_{\sigma,i}$ *are non-decreasing, if* $i \mapsto \eta_i > 0$ *is non-decreasing, and we either:*

(i) (Constant rule) Take $\eta^\perp_{\tau,i} \equiv \eta^\perp_\tau$ *and* $\eta^\perp_{\sigma,i} \equiv \eta^\perp_\sigma$ *for constant* $\eta^\perp_\sigma, \eta^\perp_\tau > 0$ *satisfying*

$$\eta_0 \cdot \min_j(\pi_j - \mathring{\pi}_j) > \eta^\perp_\tau, \quad \text{and} \quad \eta_0 \cdot \min_\ell(\nu_\ell - \mathring{\nu}_\ell) > \eta^\perp_\sigma. \tag{4.10}$$

(ii) (Proportional rule) For some $\alpha \in (0,1)$ *let us take* $\eta^\perp_{\tau,i} := \eta^\perp_{\sigma,i} := \alpha\eta_i$ *satisfying*

$$\min_j\,(\pi_j - \mathring{\pi}_j) > \alpha, \quad \text{and} \quad \min_\ell\,(\nu_\ell - \mathring{\nu}_\ell) \geq \alpha. \tag{4.11}$$

Proof. Clearly both rules satisfy (C-$\eta^\perp$), while (4.10) or (4.11) together with (R-$\pi\nu$) and $i \mapsto \eta_i$ being positive and non-decreasing, guarantee (C-η). That $i \mapsto \eta^\perp_{\tau,i}$ and $i \mapsto \eta^\perp_{\sigma,i}$ are non-decreasing is obvious. ◻

4.3 WORST-CASE RULES FOR η_i

For a random variable $p \in \mathcal{R}(\Omega;\mathbb{R})$ on the probability space $(\Omega, \mathcal{O}, \mathbb{P})$, let us define the conditional worst-case realisation with respect to the σ-algebra $\mathcal{O}' \subset \mathcal{O}$ as the random variable $\mathbb{W}[p|\mathcal{O}'] \in \mathcal{R}(\mathcal{O}';\mathbb{R})$ defined by

$$p \leq \mathbb{W}[p|\mathcal{O}'] \leq q \quad \mathbb{P}\text{-a.e.} \quad \text{for all} \quad q \in \mathcal{R}(\mathcal{O}';\mathbb{R}) \quad \text{s.t.} \quad p \leq q \quad \mathbb{P}\text{-a.e.}$$

We also write $\mathbb{W}[p] := \mathbb{W}[p|\mathcal{O}']$ when $\mathcal{O}' = \{\Omega, \emptyset\}$ is the trivial σ-algebra.

By (R-λ), (R-$\tau\sigma$), and (R-$\pi\nu$), we have

$$\lambda_{l,j,i} \leq \eta_i\left(\mathring{\pi}_j^{-1}\chi_{\mathring{S}(i)}(j) + \mathring{\nu}_\ell^{-1}\chi_{\mathring{V}(i+1)}(\ell)\right) =: \eta_i\hat{\mu}_{\ell,j,i}, \quad (\ell \in \mathcal{V}(j)).$$

The condition (C-$\kappa\psi$) will therefore hold if

$$\psi_{\ell,i+1} \geq \frac{\eta_i^2}{1-\delta}\mathbb{W}[\kappa_\ell(\ldots,\hat{\mu}^2_{\ell,j,i}\phi^{-1}_{j,i},\ldots)|\mathcal{O}_{i-1}]. \tag{4.12}$$

Accordingly, we take

$$\eta_i := \min_{\ell=1,\ldots,n}\sqrt{\frac{(1-\delta)\psi_{\ell,i+1}}{\mathbb{W}[\kappa_\ell(\ldots,\hat{\mu}^2_{\ell,j,i}\phi^{-1}_{j,i},\ldots)|\mathcal{O}_{i-1}]}}. \tag{R-η}$$

By the construction of $\mathbb{W}$, we get $\eta_i \in \mathcal{R}(\mathcal{O}_{i-1};(0,\infty))$ provided also $\psi_{i+1} \in \mathcal{R}(\mathcal{O}_{i-1};(0,\infty))$. It is our task in the rest of this section to experiment with different choices of $\psi_{\ell,i+1}$, satisfying (C-ψinc) and (C-ybnd). Before this we establish the following important fact.

Lemma 4.3. *Let $\delta \in (0,1)$ and $(\kappa_1, \dots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, Q)$. Suppose (R-$\pi\nu$) holds, and that both $i \mapsto \phi_{j,i}$ and $i \mapsto \psi_{\ell,i}$ are non-decreasing for all $j = 1, \dots, m$ and $\ell = 1, \dots, n$. Then $i \mapsto \eta_i$ defined in (R-η) is non-decreasing and ensures (C-$\kappa\psi$).*

Proof. We fix $\ell \in \{1, \dots, n\}$. The condition (R-$\pi\nu$) implies that $(\hat\mu_{\ell,1,i}, \dots, \hat\mu_{\ell,m,i})$ are independently identically distributed for all $i \in \mathbb{N}$. Since $\phi_{j,i} \in \mathcal{R}(O_{i-1}; (0,\infty))$, we can for some random $(\hat\mu_1, \dots, \hat\mu_m)$ on a probability space $(\mathbb{P}_\mu, \Omega_\mu, O_\mu)$, distinct from $(\mathbb{P}, \Omega, O)$, write

$$\mathbb{W}[\kappa_\ell(\dots, \hat\mu_{\ell,j,i}^2 \phi_{j,i}^{-1}, \dots)|O_{i-1}] \sim \mathbb{W}[\kappa_\ell(\dots, \hat\mu_j^2 \phi_{j,i}^{-1}, \dots)], \quad (i \in \mathbb{N}),$$

where $\sim$ stands for “identically distributed”. Since $i \mapsto \phi_{j,i}$ is non-decreasing and κ_ℓ monotone, this implies

$$\mathbb{W}[\kappa_\ell(\dots, \hat\mu_{\ell,j,i}^2 \phi_{j,i}^{-1}, \dots)|O_{i-1}] \geq \mathbb{W}[\kappa_\ell(\dots, \hat\mu_{\ell,j,i+1}^2 \phi_{j,i+1}^{-1}, \dots)|O_i], \quad \mathbb{P}\text{-a.e.}$$

Since $i \mapsto \psi_{\ell,i}$ is also non-decreasing, the claim follows. □

4.4 MIXED RATES UNDER PARTIAL STRONG CONVEXITY

Let us take $\psi_{\ell,i+1} := \psi_{\ell,0}\eta_i^{2-1/p}$ for some $p \in (0,1]$. Then (R-η) gives

$$\text{(R-}\eta'\text{)} \qquad \eta_i = \min_{\ell=1,\dots,n} \left(\frac{(1-\delta)\psi_{\ell,0}}{\mathbb{W}[\kappa_\ell(\dots, \hat\mu_{\ell,j,i}^2 \phi_{j,i}^{-1}, \dots)|O_{i-1}]} \right)^p.$$

Lemma 4.4. *Let $(\kappa_1, \dots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, Q)$ Suppose (R-η) and (4.1c) hold, $\phi_{j,i} \in \mathcal{R}(O_{i-1}; (0,\infty))$, and that $\psi_{\ell,i+1} = \eta_i^{2-1/p}\psi_{\ell,0}$ for some $\psi_{\ell,0} > 0$ and $p \in (0,1]$. 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. With $\underline{\psi}_0 := \min_{\ell=1,\dots,n} \psi_{\ell,0}$, from (R-$\eta$) now

$$\eta_i^{1/p} \geq \frac{(1-\delta)\underline{\psi}_0}{\max_{\ell=1,\dots,n} \mathbb{W}[\kappa_\ell(\dots, \hat\mu_{\ell,j,i}^2 \phi_{j,i}^{-1}, \dots)|O_{i-1}]}.$$

Since $\hat\mu_{\ell,j,i} = 0$ for $\ell \notin \mathcal{V}(j)$, using (C-$\kappa$.b) and $\phi_{j,i} \in \mathcal{R}(O_{i-1}; (0,\infty))$, we get

$$\eta_i^{1/p} \geq \frac{(1-\delta)\underline{\psi}_0}{\overline{\kappa} \max_{\ell \in \mathcal{V}(j)} \mathbb{W}[\sum_{j=1}^n \hat\mu_{\ell,j,i}^2 \phi_{j,i}^{-1}|O_{i-1}]} \geq \frac{1}{\sum_{j=1}^n b_j^{-1} \phi_{j,i}^{-1}}$$

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

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

Using (4.1c) follows $\mathbb{E}[\eta_i] \geq c_\eta^p i^p$ for $c_\eta := 1/\sum_{j=1}^m b_j^{-1} c_j$. □

We introduce the short-hand notation

$$\mathbb{w}_j := \max_{\ell \in \mathcal{V}(j)} \mathbb{W}[\hat{\mu}_{\ell,j,i} | O_{i-1}] = \max_{\ell \in \mathcal{V}(j)} \mathbb{W}\left[\mathring{\pi}_j^{-1} \chi_{\mathring{S}(i)}(j) + \mathring{\nu}_\ell^{-1} \chi_{\mathring{V}(i+1)}(\ell) \;\middle|\; O_{i-1}\right],$$

observing that $\mathbb{w}_j$ is independent of $i \geq 0$ by the proof of Lemma 4.3. With this, we are finally ready to state our main result:

Theorem 4.5. *Let $\delta \in (0,1)$ and $(\kappa_1, \ldots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$ (see Definition 3.1). Write $\gamma_j \geq 0$ for the factor of (strong) convexity of G_j. In Algorithm 1 or Algorithm 2, take*

(i) The probabilities $\mathbb{P}[j \in \mathring{S}(i)] \equiv \mathring{\pi}_j > 0$, and (in Algorithm 1) $\mathbb{P}[\ell \in \mathring{V}(i+1)] \equiv \mathring{\nu}_\ell$ independent of iteration, satisfying (C-SV).

(ii) $\phi_{j,0} > 0$ freely, and $\phi_{j,i+1} := \phi_{j,i} + 2(\bar{\gamma}_j \eta_i + \rho_j)$, where $\bar{\gamma}_j, \rho_j \geq 0$ satisfy $\rho_j + \bar{\gamma}_j > 0$.

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

(iv) η_i by (R-η') and (in Algorithm 1) $\eta_{\tau,i}^\perp, \eta_{\sigma,i}^\perp > 0$ following Lemma 4.2.

Let $\ell^*(j)$ and $\underline{\kappa}$ be given by (C-κ.c). Suppose for each $j = 1, \ldots, m$ either $\rho_j = 0$ or (C-xbnd.a) holds with the constant C_x, and for some $\widetilde{\gamma}_j \in [\bar{\gamma}_j, \gamma_j]$ the initialisation bound holds

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

If $p \neq 1/2$, also assume that $\bar{\gamma}_{j^*} = 0$ for some $j^* \in \{1, \ldots, m\}$, and that (C-ybnd.a) holds with the corresponding constant C_y. Let $\widetilde{c}_j \geq 0$ be the constants provided by Lemma 4.1. Then

$$\sum_{j=1}^m \delta \widetilde{c}_j \bar{\gamma}_j \mathbb{E}\left[\|x_j^N - \widehat{x}_j\|\right]^2 + g_{p,N} \leq \frac{C_0 + 18 C_x (\sum_{j=1}^m \rho_j) N + \sum_{\ell=1}^n \psi_{\ell,0} (C^* N + \delta^*)}{N^{p+1}}, \tag{4.13}$$

for $N \geq 4$, and

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

The constants $c_p, c_{*,p} > 0$ are dependent on p alone, while $C^*, \delta^* \geq 0$ are zero if $p = 1/2$, and otherwise depend on $\psi_{\ell^*(j^*),0}$, $\phi_{j^*,0}$, $\underline{\kappa}$, $\mathbb{w}_{\ell^*(j^*)}$, δ, and p.

Remark 4.6. *If $p = 1/2$, we have $\psi_{\ell,i+1} = \psi_{\ell,0}$, and a mixed $O(1/N^{3/2}) + O(1/N^{1/2})$ convergence rate. If $p = 1$, $\psi_{\ell,i+1}$ is increasing, but we have a mixed $O(1/N^2) + O(1/N)$ convergence rate.*

Remark 4.7. *The lemma is valid (with suitable constants) for general primal update rules as long as (4.1) holds and $i \mapsto \phi_{j,i}$ is non-decreasing. As we have seen, this is the case for the deterministic rule of Lemma 4.1. For the random rule of Example 3.3, the rest of the conditions hold, but we have not been able to verify (4.1e). This has the implication that only the gap estimates hold.*

Proof. We use Corollary 3.9 (Algorithm 1) and 3.10 (Algorithm 2). We have assumed (C-SV). The conditions (C-ψinc) and (C-ybnd.b) clearly hold by the choice of $\psi_{\ell,i+1}$. Since $i \mapsto \phi_{j,i}$ and $i \mapsto \psi_{\ell,i}$ are clearly non-decreasing, Lemma 4.3 shows (C-$\kappa\psi$) and that $i \mapsto \eta_i$ is non-decreasing. Moreover, (i) verifies (R-$\pi\nu$). Therefore, Lemma 4.2 shows (C-$\eta^\perp$), (C-η), and that $i \mapsto \eta^\perp_{\tau,i}$ is non-decreasing (for Algorithm 1).

To verify (C-xbnd) via Lemma 4.1, we still need to satisfy (C-ϕdet). With the help (C-κ.c), we deduce from (R-η') that

$$\eta_i \le \left(\frac{(1-\delta)\psi_{\ell^*(j),0}}{\underline{\kappa}\mathbb{W}_j}\phi_{j,i}\right)^p. \tag{4.14}$$

Moreover $\phi_{j,i}^{1-p} \ge \phi_{j,0}^{1-p}$. Therefore (C-$\phi$det) follows from (C-$\phi$det$'$). Lemma 4.1 thus shows (C-xbnd), since the algorithms satisfy (3.11). This finishes the verification of the conditions of the corollaries, so we obtain the estimate (3.17).

To obtain convergence rates, we need to further analyse (3.17). We mainly need to estimate ζ_N and $\zeta_{*,N}$. We recall the variable $\bar\eta_i$ from (C$\mathcal{G}$) and (C$\mathcal{G}_*$). The condition (C-η) and the update formulas (R-$\tau\sigma$) guarantee $\bar\eta_i \ge \mathbb{E}[\eta_i]$; cf. Section 3.4. Moreover, Lemma 4.4 gives $\mathbb{E}[\eta_i] \ge c_\eta^p i^p$ for some $c_\eta > 0$. Thus we estimate ζ_N from (2.9) as

$$\begin{aligned}\zeta_N &= \sum_{i=0}^{N-1}\bar\eta_i \ge \sum_{i=0}^{N-1}\mathbb{E}[\eta_i] \ge c_\eta^p\sum_{i=0}^{N-1} i^p \ge c_\eta^p\int_0^{N-2} x^p\,dx\\ &\ge \frac{c_\eta^p}{p+1}(N-2)^{p+1} \ge \frac{c_\eta^p}{2^{p+1}(p+1)}N^{p+1} =: c_pN^{p+1} \quad (N \ge 4),\end{aligned} \tag{4.15}$$

Similarly, for some $c_{*,p} > 0$, $\zeta_{*,N}$ defined in (2.9) satisfies

$$\zeta_{*,N} \ge \sum_{i=1}^{N-1}\mathbb{E}[\eta_i] \ge \frac{c_\eta^p}{p+1}((N-2)^{p+1}-1) \ge c_{*,p}N^{p+1} \quad (N \ge 4). \tag{4.16}$$

If $p = 1/2$, clearly $d^y_{\ell,N} = \mathbb{E}[\psi_{\ell,N} - \psi_{\ell,0}] \equiv 0$. Otherwise, we still need to bound $\psi_{\ell,N+1}$ to bound $d^y_{\ell,N}$. To do this, we need the assumed existence j^* with $\gamma_{j^*} = 0$. From (4.14) we have $\eta_i \le C\phi_{j^*,i}$ for some $C > 0$. Since $\gamma_{j^*} = 0$, a referral to (4.1b) shows that $\mathbb{E}[\phi_{j^*,N}] = \phi_{j^*,0} + N\rho_{j^*}$. We now deduce for some $C_*, \delta_* \ge 0$ that

$$\mathbb{E}[d^N_{y,\ell}] = \psi_{\ell,0}(\mathbb{E}[\eta_N] - 1) \le \psi_{\ell,0}\big(C^*N^p + \delta^*\big). \tag{4.17}$$

Lemma 4.4 shows $\eta_i \ge b_\eta^p \min_j \phi_{j,i}^p$ for $j = 1, \dots, m$. Thus (4.1e) and (4.1d) in Lemma 4.1 give $1/\mathbb{E}[\phi_{j,N}^{-1}] \ge \bar\gamma_j\widetilde{c}_jN^{p+1}$ for $N \ge 4$, and $d^x_{j,N}(\widetilde{\gamma}_j) = 18\rho_jC_xN$. Now (4.13) is immediate from applying these estimates and (4.15)–(4.17) in (3.17). $\square$

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 (R-$\eta$) shows that $\eta_i \equiv \eta_0$. Recalling ζ_N from (2.9), we see that $\zeta_N = N\eta_0$. Likewise $\zeta_{*,N}$ from (2.9) 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 (3.17), we immediately obtain:

Corollary 4.8. *Let $\delta \in (0,1)$ and $(\kappa_1, \ldots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$. In Algorithm 1 or 2, take*

(i) $\mathbb{P}[j \in \mathring{S}(i)] \equiv \mathring{\pi}_j > 0$, and (in Algorithm 1) $\mathbb{P}[\ell \in \mathring{V}(i+1)] \equiv \mathring{\nu}_\ell$ independent of iteration, satisfying (C-SV).

(ii) $\phi_{j,i} \equiv \phi_{j,0}$ for some fixed $\phi_{j,0} > 0$.

(iii) $\psi_{\ell,i} \equiv \psi_{\ell,0}$ for some fixed $\psi_{\ell,0} > 0$.

(iv) $\eta_i \equiv \eta_0$ with η_0 given by (R-η), 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 $\mathcal{G}(\widetilde{x}_N, \widetilde{y}_N) \leq C_0\eta_0^{-1}/N$, $(N \geq 1)$.

(II) The iterates of Algorithm 2 satisfy $\mathcal{G}(\widetilde{x}_{*,N}, \widetilde{y}_{*,N}) \leq C_0\eta_0^{-1}/(N-1)$, $(N \geq 2)$.

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. *Let $\delta \in (0,1)$ and $(\kappa_1, \ldots, \kappa_n) \in \mathcal{K}(K, \mathcal{P}, \mathcal{Q})$. Write γ_j for the factor of strong convexity of G_j, and suppose $\min_j \gamma_j > 0$. In Algorithm 1 or Algorithm 2, take*

(i) $\mathbb{P}[j \in \mathring{S}(i)] \equiv \mathring{\pi}_j > 0$, and (in Algorithm 1) $\mathbb{P}[\ell \in \mathring{V}(i+1)] \equiv \mathring{\nu}_\ell$ independent of iteration, satisfying (C-SV).

(ii) η_i according to (R-η), and (in Algorithm 1) $\eta^\perp_{\tau,i}, \eta^\perp_{\sigma,i} > 0$ following Lemma 4.2.

(iii) $\phi_{j,0} > 0$ freely, and $\phi_{j,i+1} := \phi_{j,i}(1 + 2\bar{\gamma}_j\tau_{j,i})$ for some fixed $\bar{\gamma}_j \in (0, \gamma_j)$.

(iv) $\psi_{\ell,i} := \psi_{\ell,0}$ for some fixed $\psi_{\ell,0} > 0$.

Suppose (C-ϕdet′) holds for some $\widetilde{\gamma}_j \in [\bar{\gamma}_j, \gamma_j]$. Let $\widetilde{c}_j$ be the constants from Lemma 4.1. Then

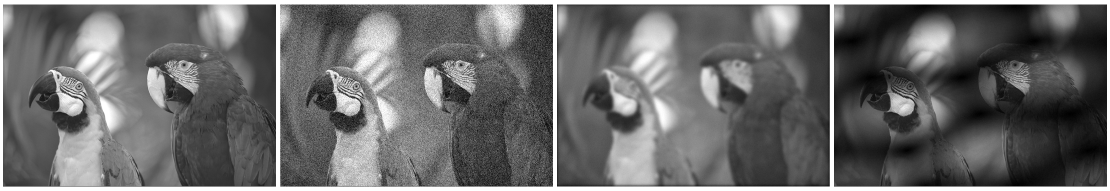
$$\sum_{j=1}^m \delta\widetilde{c}_j\bar{\gamma}_j\mathbb{E}\left[\|x_j^N - \widehat{x}_j\|\right]^2 + \widetilde{g}_{1,N} \leq \frac{C_0}{N^2}, \quad (N \geq 4).$$

where for some constants $q_1, q_{*,1} > 0$ we have

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

Proof. The update rule (R-ϕdet) now gives

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



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

Figure 1: Sample images for denoising, deblurring, and undimming experiments.

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

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

This implies $\eta_i \geq \widetilde{\eta}_i \geq c'_\eta i$ for some $c'_\eta > 0$; cf. the estimates for the acceleration rule (2.3) in [6, 33]. We now work through the proof of Theorem 4.5 with $p = 1/2$ and $\rho_j = 0$, but using in (4.15) and (4.16) the estimate $\eta_i \geq c'_\eta i$ that would otherwise correspond to $p = 1$. □

Remark 4.10 (Linear rates under full primal-dual strong convexity). *If both G and F^* are strongly convex, fixing $\tau_{j,i} \equiv \tau_j$, it is possible to derive linear rates.*

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 1, with values in the range $[0, 255]$. We use the images both at the original resolution of 768×512, and scaled down to 192×128 pixels. To the noisy high-resolution test image in Figure 1b, we have added Gaussian noise with standard deviation 29.6 (12dB). In the downscaled image, this becomes 6.15 (25.7dB). The image in Figure 1c we have distorted with 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 1d, we have distorted by multiplying the image with a sinusoidal mask γ; see Figure 1c. Again, we have added the small amount of noise to the blurry image.

Besides the unaccelerated PDHGM—our examples lack strong convexity for acceleration of basic methods—we evaluate our algorithms against the relaxed PDHGM of [7, 17]. In our precursor work [33], we have also evaluated these two algorithms against the mixed-rate method of [8], and the adaptive PDHGM of [16]. To keep our tables and figures easily legible, we also do not include the algorithms of [33] in our evaluations. It is worth noting that even in the

Table 1: Algorithm variant name construction

Letter:	1st	2nd	3rd	4th
	Randomisation	ϕ rule	η and ψ rules	κ choice
A-	D: Deterministic	R: Random, Lem. 3.3	B: Bounded: $p = \frac{1}{2}$	O: Balanc., Ex. 3.2
	P: Primal only	D: Determ., Lem. 4.1	I: Increasing: $p = 1$	M: Max., Ex. 3.1
	B: Primal & Dual	C: Constant		

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 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 (C-ybnd.a) and (C-xbnd.a), or take $\rho_j = 0$. The bound C_y in (C-ybnd) 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., [32, Appendix A]. We can therefore add an artificial bound to the problem to force all iterates to be bounded, replacing G by $\widetilde{G}(x) := G(x)+\delta_{B(0,C_x)}(x)$. In practise, 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 [32] 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 $10\log_{10}(\mathrm{gap}^2/\mathrm{gap}_0^2)$ 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 $\mathrm{val}(x) := G(x) + F(Kx)$ relative to the target as $10\log_{10}((\mathrm{val}(x) - \mathrm{val}(\hat{x}))^2/\mathrm{val}(\hat{x})^2)$. The target solution $\widehat{x}$ we compute by taking one million iterations of the basic PDHGM. Our computations were performed in Matlab+C-MEX on a MacBook Pro with 16GB RAM and a 2.8 GHz Intel Core i5 CPU. Our initial iterates are always $x^0 = 0$ 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

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

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 $P_1 x = v$, $P_2 x = w$, $Q_1 y = \phi$,

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

low resolution / 0-init						high resolution / 0-init					
gap ≤ -60dB		tgt ≤ -60dB		val ≤ -60dB		gap ≤ -50dB		tgt ≤ -50dB		val ≤ -50dB	
iter	time	iter	time	iter	time	iter	time	iter	time	iter	time
30	0.21s	100	0.72s	110	0.79s	50	6.31s	870	111.83s	370	47.49s
20	0.20s	70	0.71s	70	0.71s	40	6.93s	580	102.89s	250	44.25s
40	0.26s	230	1.55s	180	1.22s	70	9.17s	2750	365.52s	1050	139.48s
80	0.54s	890	6.07s	500	3.41s	80	10.56s	860	114.81s	420	56.00s
20	0.14s	50	0.36s	110	0.80s	60	7.37s	2140	267.29s	900	112.34s
30	0.19s	50	0.32s	90	0.58s	60	7.85s	600	79.67s	340	45.09s

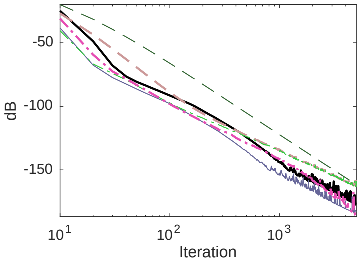
and $Q_2 y = \psi$. It is then not difficult to calculate the optimal κ_ℓ of Example 3.2, so we use only the 'xxxO' variants of the algorithms in Table 1.

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 [12]. 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.

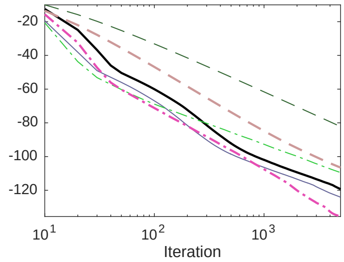
We take $\delta = 0.01$, and parametrise the standard PDHGM with $\sigma_0 = 1.9/\|K\|$ and $\tau_0 \approx 0.52/\|K\|$ solved from $\tau_0\sigma_0 = (1-\delta)\|K\|^2$. These are values that typically work well. For forward-differences discretisation of TGV^2 with cell width $h = 1$, we have $\|K\|^2 \leq 11.4$ [32]. 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.

The results for the deterministic variants of our algorithms are in Table 2 and Figure 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 slowing down the computer, the CPU times reported are based on the average iteration_time = total_time/total_iterations, excluding time spent initialising the algorithm.

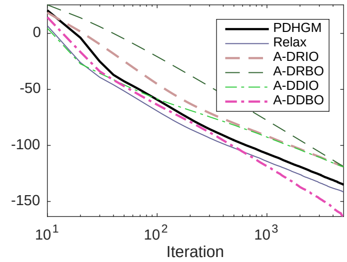
Our first observation is that the variants 'xDxx' based on the deterministic ϕ rule perform better than the "random" ϕ rule 'xRxx'. Presently, with no randomisation, the only difference between the rules is the value of $\bar{\gamma}$. The value 0.0105 from (C-ϕdet′) for $p = 1/2$ and the value 0.0090 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 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.



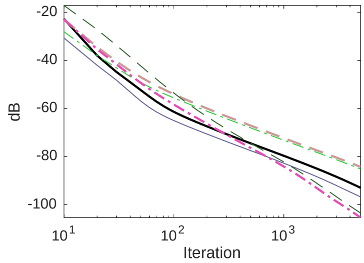
(a) Gap: lo-res



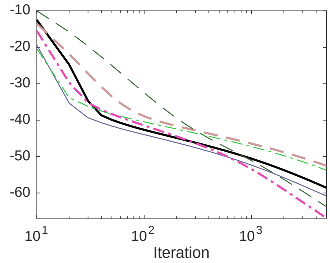
(b) Target: lo-res



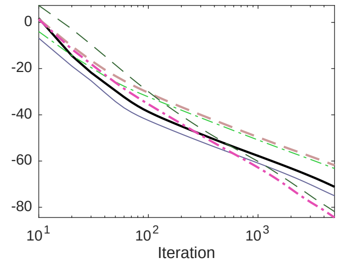
(c) Value: lo-res



(d) Gap: hi-res



(e) Target: hi-res



(f) Value: hi-res

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

5.2 TGV^2 denoising with stochastic algorithm variants

We also test stochastic variants of our algorithms based on the alternating sampling of Example 3.5 with $M = 1$ and, when appropriate, Example 3.6. We take all probabilities equal to 0.5, that is $\mathbb{p}_x = \widetilde{\pi}_1 = \widetilde{\pi}_2 = \widetilde{\nu}_1 = \widetilde{\nu}_2 = 0.5$. In the doubly-stochastic 'Bxxx' variants of the algorithms, we take $\eta^{\perp}_{\tau,i} = \eta^{\perp}_{\sigma,i} = 0.9 \cdot 0.5\eta_i$ following the proportional rule Lemma 4.2(ii).

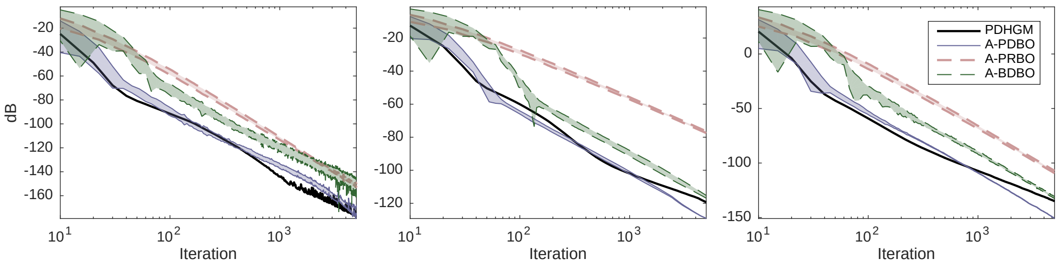
The results are in Figure 3. To conserve space, we have only included a few descriptive algorithm variants. On the x axis, to better describe to the amount of actual work performed by the stochastic methods, the "iteration" count refers to the *expected* number of full primal-dual updates. For all the 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 performance comparable to or even better than the PDHGM.

5.3 TV deblurring

We want to remove the blur in Figure 1c. We do this by taking

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



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

Figure 3: 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.

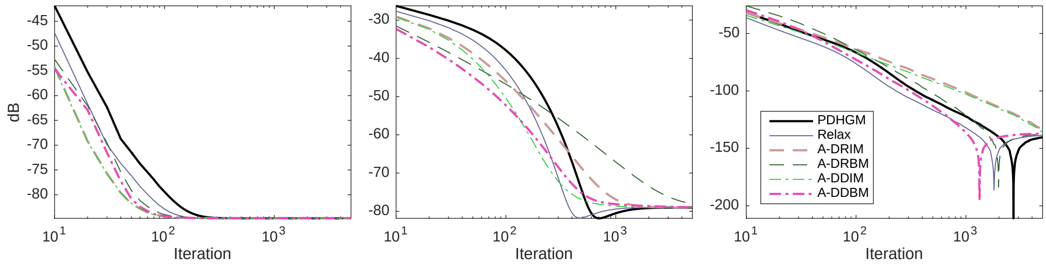
Table 3: TV deblurring performance: CPU time and number of iterations (at a resolution of 10) to reach given duality gap, distance to target, or primal objective value.

low resolution / 0-init						high resolution / 0-init					
gap ≤ -60dB		tgt ≤ -60dB		val ≤ -60dB		gap ≤ -50dB		tgt ≤ -40dB		val ≤ -40dB	
iter	time	iter	time	iter	time	iter	time	iter	time	iter	time
30	0.18s	330	2.05s	70	0.43s	60	5.04s	330	28.12s	110	9.31s
20	0.11s	220	1.30s	50	0.29s	50	4.32s	220	19.30s	90	7.84s
20	0.14s	280	2.08s	80	0.59s	30	3.27s	280	31.41s	320	35.92s
20	0.14s	490	3.58s	90	0.65s	60	6.48s	240	26.27s	220	24.07s
20	0.14s	170	1.25s	70	0.51s	30	3.17s	260	28.35s	230	25.06s
20	0.15s	180	1.37s	60	0.45s	50	5.56s	230	25.98s	150	16.90s

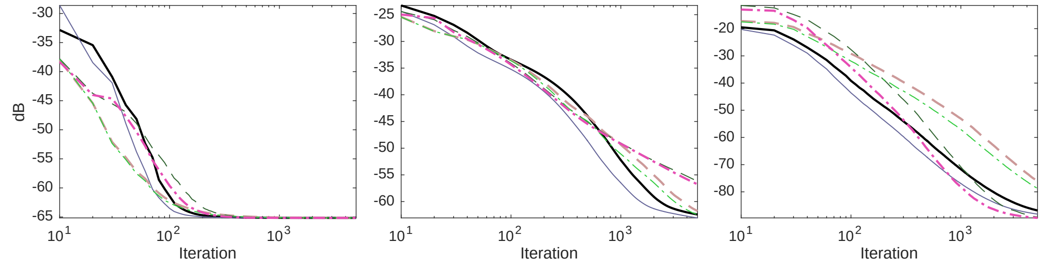
where the balls are again pixelwise Euclidean, and $\mathcal{F}$ the discrete Fourier transform. The factors $a = (a_1, \ldots, a_m)$ model the blurring operation in Fourier basis.

We use TV parameter $\alpha = 2.55$ for the high resolution image and the scaled parameter $\alpha = 2.55 * 0.15$ 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 as P_j the projection to the j:th Fourier component, and as Q_ℓ the projection to the ℓ:th pixel. Thus *each dual pixel and each primal Fourier component have their own step length parameter.* We initialise the latter as $\tau_{j,0} = \tau_0/(\lambda + (1-\lambda)\gamma_j)$, where 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 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 very difficult to compute. The results are in Table 3 and Figure 4. 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.



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



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

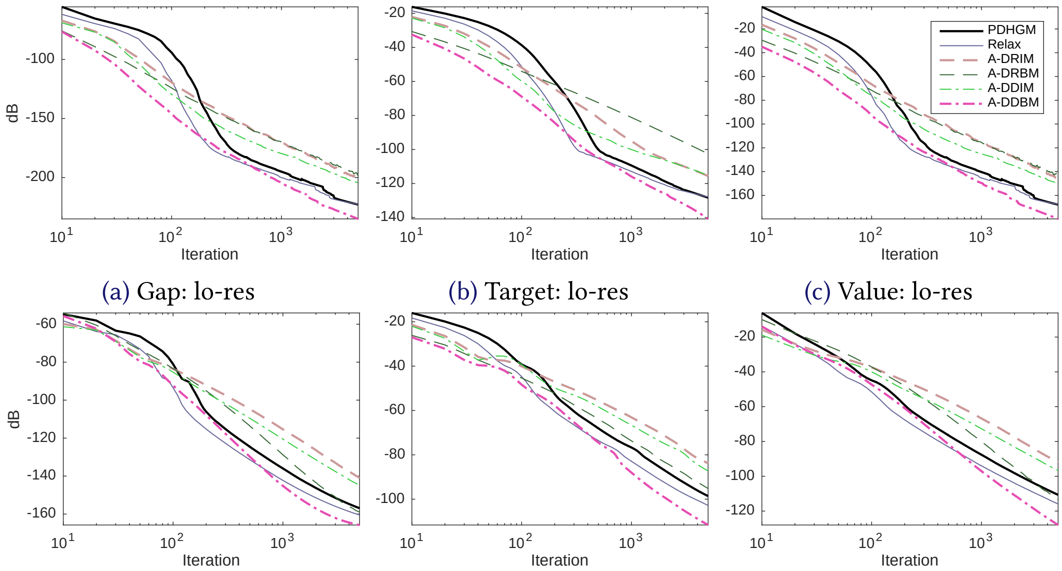
Figure 4: TV deblurring, deterministic variants of our algorithms with pixelwise step lengths, first 5000 iterations, high (hi-res) and low (lo-res) resolution images.

Table 4: TV undimming performance: CPU time and number of iterations (at a resolution of 10) to reach given duality gap, distance to target, or primal objective value.

low resolution / 0-init						high resolution / 0-init					
gap ≤ −80dB		tgt ≤ −60dB		val ≤ −60dB		gap ≤ −80dB		tgt ≤ −60dB		val ≤ −60dB	
iter	time	iter	time	iter	time	iter	time	iter	time	iter	time
70	0.18s	200	0.51s	120	0.30s	100	3.41s	300	10.31s	210	7.21s
50	0.16s	130	0.41s	80	0.25s	70	3.03s	200	8.73s	140	6.10s
30	0.10s	160	0.57s	80	0.28s	80	3.52s	760	33.82s	640	28.48s
20	0.05s	170	0.47s	60	0.16s	90	3.95s	370	16.39s	380	16.84s
30	0.08s	110	0.30s	60	0.16s	70	3.05s	580	25.57s	430	18.94s
20	0.05s	70	0.18s	40	0.10s	60	2.63s	230	10.22s	200	8.88s

5.4 TV undimming

In this problem K and F^* are as in 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 TV deblurring, with the natural difference that the projection P_j are no longer to the Fourier basis, but to individual image pixels. The results are in Figure 5, and Table 4. They tell roughly the same story as TV deblurring, with A-DDBM performing well and reliably.



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

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

Figure 5: TV undimming, deterministic variants of our algorithms with pixelwise step lengths, 5000 iterations, high (hi-res) and low (lo-res) resolution images.

CONCLUSIONS

We have derived from abstract theory several accelerated block-proximal primal-dual methods, both stochastic and deterministic. So far, we have primarily 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.

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

The codes will be archived once the final version of the paper is submitted. The sample photo is from the free Kodak image suite, at the time of writing available at http://rok.us/graphics/kodak/.

REFERENCES

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] Dimitri P Bertsekas. Incremental aggregated proximal and augmented lagrangian algorithms. 2015. preprint.

[3] Pascal Bianchi, Walid Hachem, and Franck Iutzeler. A stochastic coordinate descent primal-dual algorithm and applications to large-scale composite optimization. Preprint.

[4] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2013.

[5] A. Chambolle. An algorithm for mean curvature motion. *Interfaces and Free Boundaries*, 6(2):195, 2004.

[6] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

[7] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, pages 1–35, 2015.

[8] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[9] Patrick L Combettes and Jean-Christophe Pesquet. Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration. 2016.

[10] Dominik Csiba, Zheng Qu, and Peter Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. Preprint.

[11] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[12] Juan Carlos de Los Reyes, Carola-Bibiane Schönlieb, and Tuomo Valkonen. Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*, 57:1–25, 2017.

[13] Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[14] Olivier Fercoq and Pascal Bianchi. A coordinate descent primal-dual algorithm with large step size and possibly non separable functions. 2015.

[15] Olivier Fercoq and Peter Richtárik. Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent. *SIAM Review*, 58(4):739–771, 2016.

[16] Tom Goldstein, Min Li, and Xiaoming Yuan. Adaptive primal-dual splitting methods for statistical learning and image processing. *Advances in Neural Information Processing Systems*, 28:2080–2088, 2015.

[17] B. He and X. Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[18] Thomas Möllenhoff, Evgeny Strekalovskiy, Michael Moeller, and Daniel Cremers. The primal-dual hybrid gradient method for semiconvex splittings. *SIAM Journal on Imaging Sciences*, 8(2):827–857, 2015.

[19] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[20] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. iPiano: Inertial proximal algorithm for non-convex optimization. preprint, 2014.

[21] Zhimin Peng, Tianyu Wu, Yangyang Xu, Ming Yan, and Wotao Yin. Coordinate friendly structures, algorithms and applications. 2016.

[22] Zhimin Peng, Yangyang Xu, Ming Yan, and Wotao Yin. ARock: an algorithmic framework for asynchronous parallel coordinate updates. UCLA CAM Report 15-37, June 2015.

[23] Jean-Christophe Pesquet and Audrey Repetti. A class of randomized primal-dual algorithms for distributed optimization. 2014.

[24] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the mumford-shah functional. In *12th IEEE Conference on Computer Vision*, pages 1133–1140, Sept 2009.

[25] Zheng Qu, Peter Richtárik, and Tong Zhang. Randomized dual coordinate ascent with arbitrary sampling. Preprint, 2014.

[26] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, pages 1–52, 2015.

[27] Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. 2013.

[28] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1):105–145, 2014.

[29] A. N. Shiri͡aev. *Probability*. Graduate Texts in Mathematics. Springer, 1996.

[30] T. Suzuki. Stochastic dual coordinate ascent with alternating direction multiplier method. Preprint, 2013.

[31] Tuomo Valkonen. Testing and non-linear preconditioning of the proximal point method, 2017. Submitted.

[32] Tuomo Valkonen, Kristian Bredies, and Florian Knoll. Total generalised variation in diffusion tensor imaging. *SIAM Journal on Imaging Sciences*, 6(1):487–525, 2013.

[33] Tuomo Valkonen and Thomas Pock. Acceleration of the PDHGM on partially strongly convex functions. *Journal of Mathematical Imaging and Vision*, 2016. to appear.

[34] StephenJ. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[35] Adams Wei Yu, Qihang Lin, and Tianbao Yang. Doubly stochastic primal-dual coordinate method for empirical risk minimization and bilinear saddle-point problem. 2015.

[36] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. 2014.

[37] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling. Preprint, 2014.