Parametrization and convergence of a primal-dual block-coordinate approach to linearly-constrained nonsmooth optimization

Olivier Bilenne

PARAMETRIZATION AND CONVERGENCE OF A PRIMAL-DUAL BLOCK-COORDINATE APPROACH TO LINEARLY-CONSTRAINED NONSMOOTH OPTIMIZATION

OLIVIER BILENNE

Abstract. This note is concerned with the problem of minimizing a separable, convex, composite (smooth and nonsmooth) function subject to linear constraints. We study a randomized block-coordinate interpretation of the Chambolle-Pock primal-dual algorithm, based on inexact proximal gradient steps. A specificity of the considered algorithm is its robustness, as it converges even in the absence of strong duality or when the linear program is inconsistent. Using matrix preconditiong, we derive tight sublinear convergence rates with and without duality assumptions and for both the convex and the strongly convex settings. Our developments are extensions and particularizations of original algorithms proposed by Malitsky (2019) and Luke and Malitsky (2018). Numerical experiments are provided for an optimal transport problem of service pricing.

1. Introduction

In an p-agent network, we consider the problem

$$\min_{x\in\mathbb{R}^n} \; g(x)+h(x) \quad \text{subject to} \quad Ax=b, \tag{P}$$

where $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, and $x=(x_1,\dots,x_p)\in\mathbb{R}^n$ contains p decision variables $x_1\in\mathbb{R}^{n_1},\dots,x_p\in\mathbb{R}^{n_p}$, each locally assigned to one of the agents ($n=\sum_{i=1}^p n_i$). The functions h and g are assumed to be additively separable for the local variables, i.e.,

$$h(x)=\sum_{i=1}^p h_i(x_i), \qquad g(x)=\sum_{i=1}^p g_i(x_i),$$

where local functions $h_1,\dots,h_p$ are smooth and differentiable, and $g_1,\dots,g_p$ are convex lower semi-continuous (possibly non-smooth) functions with easily computable resolvents. We call $G(x)=\sum_{i=1}^p G_i(x_i)$ the total objective function, where $G_i(x_i)=g_i(x_i)+h_i(x_i)$, and we give A the block matrix structure $A=(A_1\,A_2\,\cdots\,A_p)$, with $A_1\in\mathbb{R}^{m\times n_1},\dots,A_p\in\mathbb{R}^{m\times n_p}$, so that the equality constraint in (P) rewrites as $\sum_{i=1}^p A_ix_i=b$. Problem (P) finds application in numerous fields, including linear programming, optimal transport, composite minimization, distributed optimization, and inverse problems.

(Olivier Bilenne) Laboratoire Informatique d'Avignon, Avignon, France
E-mail address: olivier.bilenne@univ-avignon.fr.

2020 *Mathematics Subject Classification.* 49M29, 65Y20, 90C25, 49Q22.

Keywords. Saddle-point problems · Primal-dual algorithms · Proximal gradient methods · Randomized coordinate methods · Optimal transport.

This research was supported by the Gaspard Monge Program for optimization, operations research and their interactions with data sciences (PGMO). It was done in part at the Department of Data Science and Knowledge Engineering, Maastricht University, Netherlands, and in part at CERMICS, École des Ponts ParisTech, France. The author is now with Laboratoire Informatique d'Avignon.

To solve (P) we consider a semi-distributed algorithm in the spirit of the primal-dual algorithm (PDA) of Chambolle and Pock (2011). Our developments build on the convergence study issued in Malitsky (2019) for the primal-dual method as applied to equality-constrained convex optimization. In the original analysis, the problem is reformulated as an instance of the more general constrained optimization problem

$$\min_{x\in\mathbb{R}^n} \; h(x)+g(x) \quad \text{subject to} \quad x\in \arg\min_{\tilde{x}} f(\tilde{x}), \tag{P'}$$

where the residual function f is given the quadratic form

$$f(x)=\frac{1}{2}\|Ax-b\|^2. \tag{1}$$

The primal-dual algorithm is then regarded as a variant of the accelerated proximal gradient method of Tseng (2008). An advantage of this viewpoint, as compared the classic primal-dual analysis, is that convergence can be guaranteed in the absence of a constraint qualification for strong duality, or even when the constraint $Ax=b$ is inconsistent and (P) admits no solution, in which case (P') remains feasible. We refer to Luke and Malitsky (2018) for a thorough discussion on the benefits of this particular approach to solving (P).

In the present work we revisit the coordinate-descent implementation of the PDA proposed in Luke and Malitsky (2018), extending the latter to random coordinate block selection, and to composite objective functions with smooth and nonsmooth parts by implementing inexact proximal gradient steps. The demand for proximal gradient methods, in particular, is high in the practical appplications where proximal steps cannot be implemented exactly.

For the case of strongly convex objectives, we derive a specific stepsize sequence that leads to accelerated convergence rates, which we obtain using suitable matrix preconditioners. To our knowledge, these faster rates were, to date, only available for the centralized implementation of the algorithm. Numerical experiments are eventually reported for a simple problem in optimal transport.

Notation. In the real space $\mathbb{R}^n$, the identity matrix is denoted by I_n, the vector of ones by $\mathbf{1}_n=(1,1,\dots,1)$, and $\operatorname{diag}(a_1,\dots,a_n)$ is the diagonal matrix with $a_1,\dots,a_n$ as diagonal entries. For $x,z\in\mathbb{R}^n$, we let $\langle x,z\rangle=\sum_{i=1}^n x_iz_i=x^\top z$ denote the Euclidean inner product of x and z, and $\|x\|=\langle x,x\rangle^{1/2}$ the Euclidean norm of x. Given a symmetric and positive definite matrix $\Lambda\in\mathbb{R}^{p\times p}$, we write $\|x\|_\Lambda=\langle \Lambda x,x\rangle^{1/2}$ for $x\in\mathbb{R}^n$. We call $\mathcal{S}$ the set of solutions of (P') and call G^* its optimal value, so that $G(x^*)=G^*$ for all $x^*\in\mathcal{S}$.

2. Algorithm and main results

Before discussing the algorithm, we formulate the assumptions that are made in the introduction on the convexity of g and the smoothness of h.

Assumption 2.1. *For $i=1,\dots,p$:*

(i) The effective domain $\operatorname{dom}(g_i)\subseteq\mathbb{R}^{n_i}$ is nonempty, closed, and convex. There exists a symmetric, positive semi-definite matrix $\Upsilon_i\in\mathbb{R}^{n_i\times n_i}$ such that $g_i(\cdot)-\frac{1}{2}\|\cdot\|^2_{\Upsilon_i}$ is convex.

(ii) Function h_i is convex differentiable, and there exists a symmetric, positive semi-definite matrix $\Lambda_i\in\mathbb{R}^{n_i\times n_i}$ such that, for every $x_i,\tilde{x}_i\in\operatorname{dom}(r_i)$,

$$0\leqslant h_i(\tilde{x}_i)-h_i(x_i)-\langle\nabla h_i(x_i),\tilde{x}_i-x_i\rangle\leqslant\frac{1}{2}\|\tilde{x}_i-x_i\|^2_{\Lambda_i}. \tag{2}$$

We write $\Upsilon = \mathrm{diag}(\Upsilon_1, \ldots, \Upsilon_p)$*, and* $\Lambda = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_p)$.

We note that in the general convex problem the matrices Υ and Λ are positive semi-definite ($\Upsilon, \Lambda \succcurlyeq 0$). Strong convexity of G implies as well the existence of a positive definite $\Upsilon \succ 0$. Under these assumptions, we consider the following iterative algorithm (Algorithm 1), in which we use

$$\mathrm{prox}_f^M(z) = (I + M^{-1}\partial f)^{-1} z = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2}\|x - z\|_M^2 \right\}$$

to denote the (scaled) proximal operator associated with a function f and a symmetric and positive definite scaling matrix M.

Algorithm 1: Primal-dual block coordinate descent

Parameters : P, T^k, σ^k
Initialization : $x^0 \in \mathbb{R}^n$, $y^0 = \sigma^0(Ax^0 - b)$, $u^0 = Ax^0 - b$
Output : x^k
for $k = 0, 1, 2, \ldots$ **do**
1 **select random block** $B^k(\omega) \subset \{1, \ldots, p\}$
for $i = 1, \ldots, p$ **do**
2 **if** $i \in B^k(\omega)$ **then** $x_i^{k+1} = \mathrm{prox}_{g_i}^{P_i T_i^k}\left(x_i^k - (P_i T_i^k)^{-1}(\nabla h_i(x_i^k) + A_i^\top y^k)\right)$
3 **else** $x_i^{k+1} = x_i^k$
4 $u^{k+1} = u^k + A(x^{k+1} - x^k)$
5 $y^{k+1} = y^k + \sigma^k AP(x^{k+1} - x^k) + \sigma^{k+1} u^{k+1}$

At each step, Algorithm 1 proceeds on Line 2 to individual proximal gradient steps for the primal vectors $x_1^k, \ldots, x_p^k$ along a randomly selected subset of coordinate direction blocks. The full dual vector y^k is then upated on Line 5, using the dual residual $u^{k+1} = Ax^{k+1} - b$ (recursively computed on Line 4) and a linear extrapolation term derived from the last two primal iterates. The random block selection routine on Line 1 consists in drawing successive subsets of coordinate block indices; this is done by simulating a Bernoulli process $B : \Omega \mapsto (2^{\{1,\ldots,p\}})^{\mathbb{N}}$ with underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where the probability measure $\mathbb{P}$ satisfies $\mathbb{P}(B^k = \emptyset) = 0$ and $\mathbb{P}(i \in B^k | B^{-k}) = \mathbb{P}(i \in B^k) = \pi_i > 0$ at each step k and for any block index $i \in \{1, \ldots, p\}$. In the algorithm, a weighting matrix $P = \mathrm{diag}(P_1, \ldots, P_p) \succ 0$ is combined with a sequence of block diagonal scaling matrices $T^k = \mathrm{diag}(T_1^k, \ldots, T_p^k)$ where, for $i = 1, \ldots, p$, $P_i := (1/\pi_i)\, I_{n_i}$ and $T_i^k \in \mathbb{R}^{n_i \times n_i}$ is positive definite. The products $P_i T_i^k$ are then used as a sequence of preconditioners for the proximal gradient steps.

It can be seen that the global implementation of Algorithm 1 obtained by setting $p = 1$ with $P = I_n$ and $\sigma \equiv \sigma$ yields the dual update rule $y^{k+1} = y^k + \sigma[A(2x^{k+1} - x^k) - b]$, in which case the algorithm then reduces to a simple application the PDA with parameter $\theta = 1$. The main differences between Algorithm 1 and the method studied in Luke and Malitsky (2018) are (i) the proximal steps are not computed exactly, (ii) the primal vectors are updated simultaneously in random subsets, and (iii) there is flexibility in the choice of the step size sequences T^k and σ^k. Also, we consider two distinct implementations of Algorithm 1, with either constant or decreasing stepsizes.

Constant stepsize. With fixed parameters $T^k \equiv T$ and $\sigma^k \equiv \sigma$, Algorithm 1 is a mere extension of Luke and Malitsky (2018) to composite objective functions with updates involving random blocks of coordinates. A suitable choice for the preconditioner T, then, is to set

$$T_i = \frac{I_{n_i}}{\tau_i} + \pi_i \Lambda_i + \sigma A_i^\top A_i \qquad i = 1, \dots, p \tag{3a}$$

for some nonegative parameters $(\tau_1, \dots, \tau_p)$ and σ satisfying

$$\operatorname{diag}\left(\frac{1}{\pi_1}\Big(\frac{I_{n_1}}{\tau_1} + \sigma A_1^\top A_1\Big), \dots, \frac{1}{\pi_p}\Big(\frac{I_{n_p}}{\tau_p} + \sigma A_p^\top A_p\Big)\right) - \sigma \Xi \succ 0, \tag{3b}$$

where we define $\Xi = (\Xi_{ij})$, with $\Xi_{ij} = \pi_{i,j} A_i^\top A_j / (\pi_i \pi_j)$ and $\pi_{i,j} = \mathbb{P}(\{\omega \in \Omega | i, j \in \omega\})$. We note that the matrix Ξ features nondiagonal terms only when the the coordinates are updated by blocks, i.e., if $\pi_{i,j} \neq 0$ for some $i \neq j$.

Theorem 2.2 characterizes the convergence of Algorithm 1 when stepsize (3) is used. Its statement involves a particular sequence $s^k = \frac{1}{k}\sum_{l=1}^k x^l$ or, equivalently,

$$s^k = (I_n - P)\left(\frac{\sum_{l=0}^{k-1} \sigma^l x^l}{\sum_{l=0}^{k-1} \sigma^l}\right) + P\left(\frac{\sum_{l=0}^{k-1} \sigma^l x^{l+1}}{\sum_{l=0}^{k-1} \sigma^l}\right), \tag{4}$$

which has the quality of a weighted time average of the primal iterates.

Theorem 2.2 (Constant stepsize)**.** *Let sequence $(x^k)_k$ be issued by Algorithm 1 with parameters $T^k \equiv T, \sigma^k \equiv \sigma$ satisfying* (3)*, and consider the sequence $(s^k)_k$ such that $s^k = \frac{1}{k}\sum_{l=1}^k x^l$.*

(i) If there exists a Lagrange multiplier for Problem (P')*, then (x^k) and (s^k) almost surely converge to a solution of* (P') *and, almost surely, $f(x^k) - f^* = o(1/k)$, $f(s^k) - f^* = O(1/k^2)$.*

(ii) If $\mathcal{S}$ is a bounded set and $h + g$ is bounded from below, then all limit points of (s^k) almost surely belong to $\mathcal{S}$ with $f(s^k) - f^* = O(1/k)$, and $\mathbb{E}[f(s^k) - f^*] = o(1/k)$.

Decreasing stepsizes. Accelerated convergence rates can be obtained when the objective function is strongly convex, i.e., if $\Upsilon = \operatorname{diag}(\Upsilon_1, \dots, \Upsilon_p) \succ 0$. In the strongly convex case, we set $T^k = T/\tau^k$, and

$$\sigma^k = \frac{\alpha}{\tau^k} - \beta, \tag{5}$$

where $\alpha = \upsigma_{\max}^{-1}(\Xi \Upsilon^{-1} P)$ and $\beta = \upsigma_{\max}(\Lambda \Upsilon^{-1} P)\,\alpha$, in which $\upsigma_{\max}(\cdot)$ denotes the spectral radius. Define

$$\kappa = \frac{\beta}{\alpha} = \upsigma_{\max}(\Lambda \Upsilon^{-1} P).$$

Given an initial $\tau^0 < 1/\kappa$, the stepsize τ^{k+1} is recursively computed from τ^k using

$$\tau^{k+1} = \max_{i \in \{1,\dots,p\}} \left\{ \frac{\frac{1}{2}\big(\frac{1}{\pi_i} - 1 - \kappa\big)(\tau^k)^2 + \sqrt{\Big(1 + \frac{1}{2}\big(\frac{1}{\pi_i} - \kappa\big)\tau^k\Big)^2 - \frac{1}{4}\big(\frac{2}{\pi_i} - 1 + 2\kappa\big)(\tau^k)^2}\ \tau^k}{1 + (\frac{1}{\pi_i} - \kappa)\tau^k - \kappa(\tau^k)^2} \right\}, \tag{6}$$

The primal stepsize sequence (6) is shown in Section 4 to have the asymptotic behavior

$$\tau^{k+1} - \tau^k + \frac{(\tau^k)^2}{2} = O\left((\tau^k)^3\right),$$

so that τ^k is $O(1/k)$ decreasing with exact local convergence rate $2/k$, and the stepsize σ^k for the dual iterates increases as $O(k)$. Alternatively, one can use a stepsize sequence given by the polynomial computation rule

$$\tau^{k+1} = \tau^k - \frac{(\tau^k)^2}{2} + \gamma\left(\frac{17}{8}\delta^2 + \frac{3}{4}\delta + \frac{1}{8} + \kappa\right)(\tau^k)^3 + \gamma\left(2\delta + \frac{1}{2}\right)\kappa\,(\tau^k)^4, \tag{7}$$

where $(\gamma - 1) > 0$ and $\delta = \max_{i\in\{1,\dots,p\}}\{|\frac{1}{\pi_i} - \kappa|\}$, which decreases as

$$\tau^{k+1} \leqslant \tau^k - \left(\frac{1-\epsilon}{2}\right)(\tau^k)^2 \tag{8}$$

on condition that τ^0 is taken small enough. An actual upper bound on the initial stepsize can be obtained by combining the conditions (C1), (C2) and (C3) derived in Section 4.

The following accelerated convergence rates for Algorithm 1 were derived by adjusting to our distributed setting the discussions of (Malitsky, 2019, Section 4.2).

Theorem 2.3 (Strong convexity)**.** *Let $\Upsilon \succ 0$, and denote by x^* the solution of* (P'). *Let $(x^k)_k$ be a sequence issued by Algorithm 1 with parameter σ^k given by* (5) *and with $T^k = P^{-2}\Upsilon/\tau^k$, where τ^k satisfies either* (6) *or* (7). *Consider the sequence $(s^k)_k$ such that $s^k = \frac{1}{k}\sum_{l=1}^k x^l$.*

(i) If there exists a Lagrange multiplier for Problem (P'), *then (x^k) almost surely converges to x^* with rate $\|x^k - x^*\|_{P^2T} = O(1/k)$ and, almost surely, $f(x^k) - f^* = o(1/k^3)$, $f(z^k) - f^* = O(1/k^4)$, $f(s^k) - f^* = O(1/k^4)$.*

(ii) Otherwise, (s^k) almost surely converges to x^* with $f(s^k) - f^* = O(1/k^2)$, and $\mathbb{E}[f(s^k) - f^*] = o(1/k^2)$.

3. Convergence analysis

3.1. **Interpretation as a proximal gradient algorithm.** The proofs for Theorems 2.2 and 2.3 rest on rewriting Algorithm 1 as the (primal-only) proximal gradient algorithm given as Algorithm 2, which shares similarities with the accelerated proximal gradient algorithm of Tseng (2008), now used with stepsize $\theta^k \equiv \sigma^k/S^k$, where $S^k = \sum_{l=0}^k \sigma^l$.

Algorithm 2: Expression of Algorithm 1 as an accelerated proximal gradient

Parameters : P, T^k, σ^k
Initialization : $x^0 = s^0 \in \mathbb{R}^n$, $S^{-1} = 0$, $S^0 = \sigma^0$
Output : x^k, s^k
for $k = 0, 1, 2, \dots$ **do**
  $z^k = \big(1/S^k\big)\big(S^{k-1}s^k + \sigma^k x^k\big)$
  select random block $B^k(\omega) \subset \{1, \dots, p\}$
  for $i = 1, \dots, p$ **do**
    if $i \in B^k(\omega)$ **then** $x_i^{k+1} = \operatorname{prox}_{g_i}^{P_iT_i^k}\big(x_i^k - (P_iT_i^k)^{-1}(\nabla h_i(x_i^k) + S^k\nabla_i f(z^k))\big)$
    else $x_i^{k+1} = x_i^k$
  $s^{k+1} = z^k + \big(\sigma^k/S^k\big)\,P\big(x^{k+1} - x^k\big)$
  $S^{k+1} = S^k + \sigma^{k+1}$

The equivalence between Algorithms 1 and 2 can be shown by introducing a dual variable $y^k \equiv S^k(Az^k - b)$. The dual variable allows us to write, on Line 3 of Algorithm 2,

$$S^k \nabla_i f(z^k) = A_i^\top S^k (Az^k - b) = A_i^\top y^k. \tag{9}$$

Line 5 also gives us

$$S^k(As^{k+1} - b) = y^k + \sigma^k AP(x^{k+1} - x^k). \tag{10}$$

Algorithm 2-Line 1 then yields the dual update rule

$$\begin{aligned} y^{k+1} &= S^{k+1}(Az^{k+1} - b) \\ &= A(S^k s^{k+1} + \sigma^{k+1} x^{k+1}) - S^{k+1} b \\ &= S^k(As^{k+1} - b) + \sigma^{k+1}(Ax^{k+1} - b) \\ &\overset{(10)}{=} y^k + \sigma^k AP(x^{k+1} - x^k) + \sigma^{k+1} u^{k+1}, \end{aligned} \tag{11}$$

where a second dual variable $u^k \equiv (Ax^k - b)$ has been introduced, which satisfies

$$u^{k+1} = u^k + A(x^{k+1} - x^k). \tag{12}$$

Algorithm 1 can eventually be recovered after inclusion of (9), (11), and (12) into Algorithm 2. Straightforward computations also lead to the expression previously given in (4) for the sequence s^k, which can be derived by induction on k. We note that the sequences s^k and z^k are bounded whenever the algorithm produces a bounded sequence x^k.

In Sections 3.2 to 3.5 we study the convergence of the sequences produced by Algorithm 2. For analysis purposes, we consider the auxiliary sequence defined by

$$\hat{x}^{k+1} = \operatorname{prox}_g^{PT^k}\left(x^k - (PT^k)^{-1}(\nabla h(x^k) + S^k \nabla f(z^k))\right), \tag{13}$$

which would coincide with iterate $\hat{x}^{k+1}$ if all coordinates were to be udpated at step k (i.e., in the event $B^k = \{1, \dots, p\}$).

3.2. **ESO for block coordinate sampling.** Let $\mathcal{F}^k := \sigma(x^0, s^0, z^0, \dots, x^k, s^k, z^k)$ denote the sigma algebra of the process history up to step k. The upcoming property for Algorithm 2 relates to the concept of *expected separable overapproximation* (ESO), as seen in Richtárik and Takáč (2014, 2016); Fercoq and Richtárik (2015).

Lemma 3.1 (Proximal step)**.** *In Algorithm 2,* $\forall x \in \mathbb{R}^n$*,*

$$g(\hat{x}^{k+1}) - g(x) \leqslant \langle \nabla\zeta(\hat{x}^{k+1}), x - \hat{x}^{k+1}\rangle \leqslant \zeta(x) - \zeta(\hat{x}^{k+1}) - \frac{1}{2}\|x - \hat{x}^{k+1}\|^2_{PT^k + \Upsilon}, \tag{14}$$

where

$$\zeta(x) = h(x^k) + \langle \nabla h(x^k), x - x^k\rangle + S^k\langle \nabla f(z^k), x - z^k\rangle + \tfrac{1}{2}\|x - x^k\|^2_{PT^k}. \tag{15}$$

Proof. Equation (13) rewrites as $\hat{x}^{k+1} = \arg\min_{\tilde{x}}\{g(\tilde{x}) + \zeta(\tilde{x})\}$. Hence, $0 \in \partial g(\hat{x}^{k+1}) + \nabla\zeta(\hat{x}^{k+1})$, which yields the first inequality. The second inequality follows by strong convexity of $g + \zeta$ with modulus $PT^k + \Upsilon$. □

3.3. **Extrapolation.** The next lemma characterizes the sequence (s^k) as a convex combination of past primal iterates. It is an extension of Lemma 2 in Fercoq and Richtárik (2015).

Lemma 3.2. *In Algorithm 2, we have*

$$s^k = \sum_{l=0}^{k} \Gamma^{k,l} x^l, \qquad k \geqslant 1, \tag{16}$$

where $(\Gamma^{k,l})$ *is a sequence of diagonal matrices such that* $\Gamma^{1,0} = I_n - P$, $\Gamma^{1,1} = P$ *and, for* $k \geqslant 1$,

$$S^k \Gamma^{k+1,l} = \begin{cases} S^{k-1}\Gamma^{k,l} & \text{for } l = 0, \dots, k-1, \\ \sigma^{k-1}P - \sigma^k(P - I_n) & \text{if } l = k, \\ \sigma^k P & \text{if } l = k+1. \end{cases} \tag{17}$$

Moreover, $\Gamma^{k+1,k} = [S^{k-1}\Gamma^{k,k} - \sigma^k(P - I_n)]/S^k$ *and* $\sum_{l=0}^{k} \Gamma^{k,l} = I_n$.

Proof. We proceed by induction. By combining Lines 1 and 5 in Algorithm 2, we find

$$s^{k+1} = \frac{1}{S^k}\left(S^{k-1}s^k + \sigma^k P x^{k+1} - \sigma^k(P - I_n)x^k\right), \tag{18}$$

which yields $s^1 = (I_n - P)x^0 + Px^1$, and the values of $\Gamma^{1,0}$ and $\Gamma^{1,1}$. Suppose now that (16) holds for $k \geqslant 1$, then we infer from (18) that

$$\begin{aligned} S^k s^{k+1} &= \textstyle\sum_{l=0}^{k} S^{k-1}\Gamma^{k,l}x^l + \sigma^k P x^{k+1} - \sigma^k(P - I_n)x^k \\ &= \textstyle\sum_{l=0}^{k-1} S^{k-1}\Gamma^{k,l}x^l + S^{k-1}\Gamma^{k,k}x^k + \sigma^k P x^{k+1} - \sigma^k(P - I_n)x^k \\ &= \textstyle\sum_{l=0}^{k-1} S^{k-1}\Gamma^{k,l}x^l + \left[S^{k-1}\Gamma^{k,k} - \sigma^k(P - I_n)\right] x^k + \sigma^k P x^{k+1}. \end{aligned} \tag{19}$$

Equation (17) follows by inspection of (16) and (19). Hence, (16) holds for all k.

It can be checked that s^k is a convex combination of $x^0, \dots, x^k$. Again, we do this by induction. We already know that $\sum_{l=0}^{1} \Gamma^{1,l} = (I_n - P) + P = I_n$ and, by (17), $\sum_{l=0}^{2} \Gamma^{2,l} = \left(1/S^1\right)\left[S^0\Gamma^{1,0} + \sigma^0 P - \sigma^1(P - I_n) + \sigma^1 P\right] = \left(1/S^1\right)\left[S^0 + \sigma^1\right] I_n = I_n$. Suppose now that $\sum_{l=0}^{k} \Gamma^{k,l} = I_n$ is true for some $k \geqslant 1$, then

$$\begin{aligned} \textstyle\sum_{l=0}^{k+1} \Gamma^{k+1,l} &\overset{(17)}{=} \left(1/S^k\right)\left[S^{k-1}\textstyle\sum_{l=0}^{k-1}\Gamma^{k,l} + \sigma^{k-1}P - \sigma^k(P - I_n) + \sigma^k P\right] \\ &= \left(1/S^k\right)\left[S^{k-1}(I_n - \Gamma^{k,k}) + \sigma^{k-1}P + \sigma^k I_n\right] \\ &\overset{(17)}{=} \left(1/S^k\right)\left[S^{k-1}\left(I_n - \left(\sigma^{k-1}/S^{k-1}\right)P\right) + \sigma^{k-1}P + \sigma^k I_n\right] = \left(1/S^k\right)\left[S^{k-1} + \sigma^k\right] I_n = I_n, \end{aligned}$$

and the last identity holds by induction on k. Observe that this last result could also be obtained immediately by simple inspection of (4). □

Now, define $\mathcal{G} = (G_1, \dots, G_p)$ and $\hat{G}^k = \langle \hat{\mathcal{G}}^k, \mathbf{1}_p \rangle$, where

$$\hat{\mathcal{G}}^k = \sum_{l=0}^{k} \Gamma^{k,l}\mathcal{G}(x^l), \qquad k \geqslant 1. \tag{20}$$

By convexity, it follows from (16) and (20) that $\hat{\mathcal{G}}^k \geqslant \mathcal{G}(s^k)$ and $\hat{G}^k \geqslant G(s^k)$. We show the following result.

Lemma 3.3. *In Algorithm 2 and for any positive definite* $M = \mathrm{diag}(M_1, \ldots, M_p)$,

$$\mathbb{E}[\|x^{k+1} - x^*\|^2_{PM}|\mathcal{F}^k] = \|\hat{x}^{k+1} - x^*\|^2_M + \|x^k - x^*\|^2_{(P-I_n)M}, \tag{21}$$

$$\mathbb{E}[\hat{G}^{k+1}|\mathcal{F}^k] = \tfrac{1}{S^k}[S^{k-1}\hat{G}^k + \sigma^k G(\hat{x}^{k+1})], \tag{22}$$

where $\hat{x}^{k+1}$ *and* $\hat{G}^k$ *are defined as in* (13) *and* (20).

Proof. For $i = 1, \ldots, p$, We find,

$$\begin{aligned}\mathbb{E}[\|x_i^{k+1} - x_i^*\|_{P_iM_i}|\mathcal{F}^k] &= \pi_i\|\hat{x}_i^{k+1} - x_i^*\|^2_{P_iM_i} + (1-\pi_i)\|x_i^k - x_i^*\|^2_{P_iM_i} \\ &= \|\hat{x}_i^{k+1} - x_i^*\|^2_{M_i} + (\pi_i^{-1} - 1)\|x_i^k - x_i^*\|^2_{M_i}.\end{aligned} \tag{23}$$

Summing up the above for $i = 1, \ldots, p$ gives (21). Next, observe that $\mathbb{E}[G_i(x^{k+1})|\mathcal{F}^k] = \pi_i G_i(\hat{x}^{k+1}) + (1-\pi_i)G_i(x^k)$ for $i \in \{1, \ldots, p\}$, which in matrix form rewrites as

$$\mathbb{E}[\mathcal{G}(x^{k+1})|\mathcal{F}^k] = P^{-1}\mathcal{G}(\hat{x}^{k+1}) + (I_n - P^{-1})\mathcal{G}(x^k). \tag{24}$$

It follows that

$$\begin{aligned}\mathbb{E}[\hat{\mathcal{G}}^{k+1}|\mathcal{F}^k] &\overset{(20)}{=} \textstyle\sum_{l=0}^{k-1} \Gamma^{k+1,l}\mathcal{G}(x^l) + \Gamma^{k+1,k}\mathcal{G}(x^k) + \Gamma^{k+1,k+1}\mathbb{E}[\mathcal{G}(x^{k+1})|\mathcal{F}^k] \\ &\overset{(17)}{=} \left(\tfrac{S^{k-1}}{S^k}\right)\textstyle\sum_{l=0}^{k-1} \Gamma^{k,l}\mathcal{G}(x^l) + \Gamma^{k+1,k}\mathcal{G}(x^k) + \left(\tfrac{\sigma^k}{S^k}\right)P\mathbb{E}[\mathcal{G}(x^{k+1})|\mathcal{F}^k] \\ &\overset{(24)}{=} \left(\tfrac{S^{k-1}}{S^k}\right)\textstyle\sum_{l=0}^{k-1} \Gamma^{k,l}\mathcal{G}(x^l) + [\Gamma^{k+1,k} + (\tfrac{\sigma^k}{S^k})(P - I_n)]\mathcal{G}(x^k) + \left(\tfrac{\sigma^k}{S^k}\right)\mathcal{G}(\hat{x}^{k+1}) \\ &\overset{(20)}{=} \left(\tfrac{S^{k-1}}{S^k}\right))\hat{\mathcal{G}}^k + \left(\tfrac{\sigma^k}{S^k}\right)\mathcal{G}(\hat{x}^{k+1}),\end{aligned} \tag{25}$$

which yields (22) since $G = \langle \mathcal{G}, \mathbf{1}_p\rangle$ and $\hat{G}^k = \langle \hat{\mathcal{G}}^k, \mathbf{1}_p\rangle$. □

3.4. **Properties of f.** In this section we report useful properties related to the penalty function f. First notice that, for all $x, \tilde{x}, \bar{x} \in \mathbb{R}^q$,

$$\begin{aligned}\langle \nabla f(x), \tilde{x} - \bar{x}\rangle &= \langle \nabla f(x), \tilde{x} - x\rangle - \langle \nabla f(x), \bar{x} - x\rangle \\ &= f(\tilde{x}) - f(x) - \tfrac{1}{2}\|A(\tilde{x} - x)\|^2 - [f(\bar{x}) - f(x) - \tfrac{1}{2}\|A(\bar{x} - x)\|^2] \\ &= f(\tilde{x}) - f(\bar{x}) - \tfrac{1}{2}\|A(\tilde{x} - x)\|^2 + \tfrac{1}{2}\|A(\bar{x} - x)\|^2.\end{aligned} \tag{26}$$

For all $x, \tilde{x} \in \mathbb{R}^q$, and $\alpha \in [0, 1]$, we also have

$$f(\alpha x + (1-\alpha)\tilde{x}) = \alpha f(x) + (1-\alpha)f(\tilde{x}) - \frac{\alpha(1-\alpha)}{2}\|A(x - \tilde{x})\|^2. \tag{27}$$

For any $x^* \in \arg\min_x f(x)$, we have $\nabla f(x^*) = 0$, and f satisfies

$$f(x) - f(x^*) = \frac{1}{2}\langle A^\top A(x - x^*), x - x^*\rangle = \frac{1}{2}\|A(x - x^*)\|^2, \quad \forall x \in \mathbb{R}^q. \tag{28}$$

Besides, we know from Line 1 in Algorithm 2 that

$$\begin{aligned}\nabla f(z^k) &= \nabla f((S^{k-1}s^k + \sigma^k x^k)/S^k) = A^\top[A(S^{k-1}s^k + \sigma^k x^k)/S^k - b] \\ &= \tfrac{1}{S^k}[S^{k-1}A^\top(As^k - b) + \sigma^k A^\top(Ax^k - b)] = \tfrac{1}{S^k}(S^{k-1}\nabla f(s^k) + \sigma^k \nabla f(x^k)),\end{aligned} \tag{29}$$

and it follows from Lines 1 and 5 in Algorithm 2 that

$$\begin{aligned}f(s^{k+1}) &= f\left(\left(\tfrac{S^{k-1}}{S^k}\right)s^k + \left(\tfrac{\sigma^k}{S^k}\right)[x^k + P(x^{k+1} - x^k)]\right) \\ &\overset{(27)}{=} \left(\tfrac{S^{k-1}}{S^k}\right)f(s^k) + \left(\tfrac{\sigma^k}{S^k}\right)f(x^k + P(x^{k+1} - x^k)) - \tfrac{1}{2}\left(\tfrac{\sigma^k S^{k-1}}{(S^k)^2}\right)\|A(x^k + P(x^{k+1} - x^k) - s^k)\|^2.\end{aligned} \tag{30}$$

The last result of this section requires new notions to be introduced. For $i \in \{1, \ldots, p\}$ let U_i be the $n \times n$ block matrix defined by $U_i = \mathrm{diag}(0, \ldots, I_{n_i}, 0, \ldots, 0)$. Clearly $\sum_{i=1}^p U_i = I_n$,

and $U_i x = (0, \ldots, x_i, \ldots, 0)$ for any $x = (x_1, \ldots, x_p) \in \mathbb{R}^n$. Then, for any random block $B \in \mathcal{I}$, we define the $n \times n$ matrix $U(B) = \sum_{i \in B} U_i$. We find $\mathbb{E}[U(B)P] = I_n$ and $\Xi = \mathbb{E}[U(B)PA^\top APU(B)]$, where the expectations are taken with respect to the probabbility measure $\mathbb{P}$. It follows that

$$
\begin{aligned}
&\frac{1}{2}\|A(\hat{x}^{k+1} - s^k)\|^2 = \frac{1}{2}\|\mathbb{E}[A(x^k + P(x^{k+1} - x^k) - s^k)|\mathcal{F}^k]\|^2 \\
&= \frac{1}{2}\mathbb{E}[\|A(x^k + P(x^{k+1} - x^k) - s^k)\|^2|\mathcal{F}^k] - \frac{1}{2}\mathbb{E}[\|A(PU(B) - I)(\hat{x}^{k+1} - x^k)\|^2|\mathcal{F}^k] \\
&= \frac{1}{2}\mathbb{E}[\|A(x^k + P(x^{k+1} - x^k) - s^k)\|^2|\mathcal{F}^k] - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{\Xi - A^\top A} \\
&\overset{(30)}{=} \Big(\frac{S^k}{\sigma^k}\Big) f(s^k) - \Big(\frac{(S^k)^2}{\sigma^k S^{k-1}}\Big)\mathbb{E}[f(s^{k+1})|\mathcal{F}^k] + \Big(\frac{S^k}{S^{k-1}}\Big)\mathbb{E}[f(x^k + P(x^{k+1} - x^k))|\mathcal{F}^k] \\
&\qquad\qquad - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{\Xi - A^\top A} \\
&\overset{(1)}{=} \Big(\frac{S^k}{\sigma^k}\Big) f(s^k) - \Big(\frac{(S^k)^2}{\sigma^k S^{k-1}}\Big)\mathbb{E}[f(s^{k+1})|\mathcal{F}^k] - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{\Xi - A^\top A} \\
&\qquad\qquad + \frac{1}{2}\Big(\frac{S^k}{S^{k-1}}\Big)\mathbb{E}[\|[A\hat{x}^{k+1} - b] + A[x^k + P(x^{k+1} - x^k) - \hat{x}^{k+1}]\|^2|\mathcal{F}^k] \\
&= \Big(\frac{S^k}{\sigma^k}\Big) f(s^k) - \Big(\frac{(S^k)^2}{\sigma^k S^{k-1}}\Big)\mathbb{E}[f(s^{k+1})|\mathcal{F}^k] - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{\Xi - A^\top A} \\
&\qquad\qquad + \Big(\frac{S^k}{S^{k-1}}\Big)\big[f(\hat{x}^{k+1}) + \frac{1}{2}\mathbb{E}[\|A(PU(B) - I)(\hat{x}^{k+1} - x^k)\|^2|\mathcal{F}^k]\big] \\
&= \Big(\frac{S^k}{\sigma^k}\Big) f(s^k) - \Big(\frac{(S^k)^2}{\sigma^k S^{k-1}}\Big)\mathbb{E}[f(s^{k+1})|\mathcal{F}^k] + \frac{1}{2}\Big(\frac{\sigma^k}{S^{k-1}}\Big)\|\hat{x}^{k+1} - x^k\|^2_{\Xi - A^\top A} + \Big(\frac{S^k}{S^{k-1}}\Big) f(\hat{x}^{k+1}),
\end{aligned}
\tag{31}
$$

where we have used $\mathbb{E}[A[x^k + P(x^{k+1} - x^k) - \hat{x}^{k+1}]|\mathcal{F}^k] = 0$.

3.5. **Main descent argument.** We now derive the main argument for the theorems. Let x^* be a solution of (P'). It follows from Lemma 3.1 that

$$
\begin{aligned}
&G(\hat{x}^{k+1}) - G^* \\
&\overset{(2)}{\leqslant} g(\hat{x}^{k+1}) + h(x^k) - G^* + \langle \nabla h(x^k), \hat{x}^{k+1} - x^k\rangle + \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{\Lambda} \\
&\overset{(15)}{=} [g(\hat{x}^{k+1}) + \zeta(\hat{x}^{k+1})] - G^* - S^k\langle \nabla f(z^k), \hat{x}^{k+1} - z^k\rangle - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k - \Lambda} \\
&\overset{(14)}{\leqslant} [g(x^*) + \zeta(x^*) - \frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{PT^k + \Upsilon}] - G^* - S^k\langle \nabla f(z^k), \hat{x}^{k+1} - z^k\rangle - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k - \Lambda} \\
&\overset{(15)}{=} h(x^k) - h(x^*) + \langle \nabla h(x^k), x^* - x^k\rangle + S^k\langle \nabla f(z^k), x^* - \hat{x}^{k+1}\rangle + \frac{1}{2}\|x^k - x^*\|^2_{PT^k} \\
&\qquad\qquad - \frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{PT^k + \Upsilon} - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k - \Lambda} \\
&\overset{(2)}{\leqslant} S^k\langle \nabla f(z^k), x^* - \hat{x}^{k+1}\rangle + \frac{1}{2}\|x^k - x^*\|^2_{PT^k} - \frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{PT^k + \Upsilon} - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k - \Lambda}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(29)}{=} S^{k-1}\langle \nabla f(s^k), x^* - \hat{x}^{k+1}\rangle + \sigma^k \langle \nabla f(x^k), x^* - \hat{x}^{k+1}\rangle + \frac{1}{2}\|x^k - x^*\|^2_{PT^k} \\
&\qquad\qquad -\frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{PT^k+\Upsilon} - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k\Lambda} \\
&\overset{(26)}{=} -S^k(f(\hat{x}^{k+1}) - f^*) - \frac{S^{k-1}}{2}\|A(x^* - s^k)\|^2 + \frac{S^{k-1}}{2}\|A(\hat{x}^{k+1} - s^k)\|^2 - \frac{\sigma^k}{2}\|A(x^* - x^k)\|^2 \\
&\qquad + \frac{\sigma^k}{2}\|A(\hat{x}^{k+1} - x^k)\|^2 + \frac{1}{2}\|x^k - x^*\|^2_{PT^k} - \frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{PT^k+\Upsilon} - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k-\Lambda} \\
&\overset{(1)}{=} -[S^k f(\hat{x}^{k+1}) + S^{k-1} f(s^k) + \sigma^k f(x^k) - 2S^k f^*] + \frac{S^{k-1}}{2}\|A(\hat{x}^{k+1} - s^k)\|^2 \\
&\qquad\qquad + \frac{1}{2}\|x^k - x^*\|^2_{PT^k} - \frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{PT^k+\Upsilon} - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k-\sigma^k A^\top A-\Lambda} \\
&\overset{(31)}{=} \frac{1}{2}\|x^k - x^*\|^2_{PT^k} + \big(\frac{(S^{k-1})^2}{\sigma^k}\big)(f(s^k) - f^*) - \frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{PT^k+\Upsilon} \\
&\qquad\qquad -\big(\frac{(S^k)^2}{\sigma^k}\big)\mathbb{E}[f(s^{k+1}) - f^*|\mathcal{F}^k] - \frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT^k-\sigma^k\Xi-\Lambda} - \sigma^k(f(x^k) - f^*).
\end{aligned}
$$

Using (21) and (22), setting $T^k = T/\tau^k$, and multiplying by σ^k, we find

$$
\begin{aligned}
\mathbb{E}\left[\tfrac{\sigma^k}{2\tau^k}\|x^{k+1} - x^*\|^2_{P^2T+\tau^k P\Upsilon} + S^k F_{k+1}|\mathcal{F}^k\right] &\leqslant \tfrac{\sigma^k}{2\tau^k}\|x^k - x^*\|^2_{P^2T+\tau^k(P-I_n)\Upsilon} + S^{k-1}F_k \\
&\quad -\tfrac{\sigma^k}{2\tau^k}\|\hat{x}^{k+1} - x^k\|^2_{PT-\tau^k(\sigma^k\Xi+\Lambda)} - (\sigma^k)^2(f(x^k) - f^*),
\end{aligned} \tag{32}
$$

where we define $F_k = \hat{G}^k - G^* + S^{k-1}(f(s^k) - f^*)$. For the convergence of the above sequence we require, at every k,

$$
\tau^k(\sigma^k\Xi + \Lambda) - PT \preccurlyeq 0, \tag{33a}
$$

$$
\frac{\sigma^{k+1}}{\tau^{k+1}}(P^2T + \tau^{k+1}(P - I_n)\Upsilon) - \frac{\sigma^k}{\tau^k}(P^2T + \tau^k P\Upsilon) \preccurlyeq 0. \tag{33b}
$$

After introducing

$$
V^k(x) = \frac{\sigma^k}{2\tau^k}\|x^k - x\|^2_{P^2T+\tau^k P\Upsilon} + S^{k-1}F_k,
$$

we can rewrite (32) as the inequality

$$
\mathbb{E}\left[V^{k+1}(x^*)|\mathcal{F}^k\right] \leqslant V^k(x^*) - \frac{\sigma^k}{2\tau^k}\|\hat{x}^{k+1} - x^k\|^2_{PT-\tau^k(\sigma^k\Xi+\Lambda)} - (\sigma^k)^2(f(x^k) - f^*)\,, \tag{34}
$$

which becomes the focal point in our convergence analysis. Moreover, taking the total expectation in (34) and iterating on k gives

$$
\mathbb{E}\left[V^k(x^*) + \sum\nolimits_{l=0}^{k-1}\frac{\sigma^l}{2\tau^l}\|\hat{x}^k - x^l\|^2_{PT-\tau^l(\sigma^l\Xi+\Lambda)}\right] \leqslant C. \tag{35}
$$

where $C := \frac{\sigma^0}{2\tau^0}\|x^0 - x^*\|^2_{P^2T+\tau^0(P-I_n)\Upsilon}$. It follows from $\hat{G}^k \geqslant G(s^k)$ and the definitions of $V^k(x^*)$ and F_k, that

$$
\mathbb{E}\left[\frac{\sigma^k}{2\tau^k S^{k-1}}\|x^k - x^*\|^2_{P^2T+\tau^k P\Upsilon}\right] + S^{k-1}\mathbb{E}[f(s^k) - f^*] + \mathbb{E}[G(s^k) - G^*] \leqslant \frac{C}{S^{k-1}}. \tag{36}
$$

Besides, if G is bounded from below by a finite constant $G_{\min}$, then we also find

$$\mathbb{E}[f(s^k) - f^*] \leqslant \frac{C}{(S^{k-1})^2} + \frac{G^* - G_{\min}}{S^{k-1}}. \tag{37}$$

Recall the sequence (34). Convergence of $(V^k(x))$ follows almost surely from (34) on condition that $V^k(x)$ is bounded from below, in which case Doob's supermartingale theorem applies. As we see below, such a lower bound exists in the case when strong duality holds in (P') with respect to the constraint $\nabla f(x) = A^\top(Ax - b) = 0$ and an optimal multiplier can be found (Section 3.6). In the negative, it is necessary to step back and to consider the convergence of a downscaled version of the sequence $V^k(x)$ (Section 3.7).

3.6. **Duality and existence of a Lagrange multiplier.** Rewrite the constraint in (P') as $\nabla f(x) = 0$, i.e., $A^\top Ax - A^\top b = 0$. First assume that (P') admits an optimal primal-dual pair (x^*, v^*) satisfying $0 \in \partial G(x^*) + A^\top Av^*$, or

$$\begin{aligned} G(s^k) - G^* = G(s^k) - G(x^*) &\geqslant \langle -A^\top Av^*, s^k - x^*\rangle = -\langle Av^*, A(s^k - x^*)\rangle \\ &\geqslant -\|Av^*\|\|A(s^k - x^*)\| \overset{(28)}{=} -\|Av^*\|\sqrt{2(f(s^k) - f^*)}, \end{aligned} \tag{38}$$

where we have used $A^\top(Ax^* - b) = 0$. Since $\hat{G}^k \geqslant G(s^k)$, we find

$$\begin{aligned} S^{k-1}F_k = S^{k-1}[\hat{G}^k - G^* + S^{k-1}(f(s^k) - f^*)] &\geqslant S^{k-1}[G(s^k) - G^* + S^{k-1}(f(s^k) - f^*)] \\ &\overset{(38)}{\geqslant} (S^{k-1})^2(f(s^k) - f^*) - \sqrt{2}\|Av^*\|\sqrt{(S^{k-1})^2(f(s^k) - f^*)} \\ &= \tfrac{1}{2}(S^{k-1})^2(f(s^k) - f^*) + \left(\sqrt{\tfrac{1}{2}(S^{k-1})^2(f(s^k) - f^*)} - \|Av^*\|\right)^2 - \|Av^*\|^2 \\ &\geqslant \frac{1}{2}(S^{k-1})^2(f(s^k) - f^*) - \|Av^*\|^2 \geqslant -\|Av^*\|^2. \end{aligned} \tag{39}$$

It follows that for all x the sequence $V^k(x) \geqslant -\|Av^*\|^2$ is bounded from below, and the supermartingale theorem applies in (34), so that the sequence $(V^k(x^*))$ converges almost surely to a random quantity not smaller than $-\|Av^*\|^2$, and

$$\sum_{k=0}^{\infty}\left[\frac{\sigma^k}{2\tau^k}\|\hat{x}^{k+1} - x^k\|^2_{PT - \tau^k(\sigma^k\Xi + \Lambda)} + (\sigma^k)^2(f(x^k) - f^*)\right] < \infty \quad \text{a.s.}. \tag{40}$$

The convergence analysis of the sequences x^k, s^k and z^k is deferred to the proofs of Theorems 2.2(i) and 2.3(i) in the Appendix.

3.7. **Without a Lagrange multiplier.** Suppose now that no such optimal primal-dual pair exists. If G is bounded from below by a finite constant $G_{\min}$, then $\hat{G}^k \geqslant G(s^k) \geqslant G_{\min}$, and we find

$$F_k = \hat{G}^k - G^* + S^{k-1}(f(s^k) - f^*) \geqslant G_{\min} - G^* + S^{k-1}(f(s^k) - f^*) \geqslant G_{\min} - G^*. \tag{41}$$

The supermartingale theorem can now be applied to the sequence defined by $W^0(x^*) = V^0(x^*) = \frac{\sigma^0}{2\tau^0}\|x^0 - x^*\|^2_{P^2T + \tau^0(P - I_n)\Upsilon}$ and $W^k(x^*) = V^k(x^*)/S^{k-1}$ for $k \geqslant 1$. This downscaled sequence is bounded from below by $G_{\min} - G^*$ and satisfies

$$\begin{aligned} \mathbb{E}\left[W^{k+1}(x^*)|\mathcal{F}^k\right] &\leqslant \frac{1}{S^{k-1}}\mathbb{E}\left[V^{k+1}(x^*)|\mathcal{F}^k\right] \\ &\overset{(34)}{\leqslant} W^k(x^*) - \frac{\sigma^k}{2\tau^k S^{k-1}}\|\hat{x}^{k+1} - x^k\|^2_{PT - \tau^k(\sigma^k\Xi + \Lambda)} - \frac{(\sigma^k)^2}{S^{k-1}}(f(x^k) - f^*), \end{aligned} \tag{42}$$

where we have used $S^k \geqslant S^{k-1}$. We find that $W^k(x^*)$ converges almost surely to a random quantity $Z(\omega) \geqslant G_{\min} - G^*$, and

$$\sum_{k=0}^{\infty} \left[\frac{\sigma^k}{2\tau^k S^{k-1}} \|\hat{x}^{k+1} - x^k\|^2_{PT-\tau^k(\sigma^k \Xi+\Lambda)} + \frac{(\sigma^k)^2}{S^{k-1}} (f(x^k) - f^*) \right] < \infty \quad \text{a.s..} \tag{43}$$

See the proofs of Theorems 2.2(ii) and 2.3(ii) in the Appendix for a derivation of the convergence rates of the individual sequences. We note that the orders of convergence lost in the absence of a Lagrange multiplier are due to the scaling factor $1/S^{k-1}$ in the sequence $W^k(x^*)$.

4. Decreasing stepsizes for the strongly convex case

In this section we derive the decreasing stepsize sequence (6) for the strongly convex case ($\Upsilon \succ 0$). First, we ensure (33a) holds by setting $\sigma^k = \alpha/\tau^k - \beta$, with $\alpha = \sigma^{-1}_{\max}(\Xi(PT)^{-1})$ and $\beta = \sigma_{\max}(\Lambda(PT)^{-1})\,\alpha$. Indeed, we find $1/\tau^k = \sigma^k/\alpha + \sigma_{\max}(\Lambda(PT)^{-1}) = \sigma^k\, \sigma_{\max}(\Xi(PT)^{-1}) + \sigma_{\max}(\Lambda(PT)^{-1})$. Consequently,

$$\begin{aligned} \tau^k(\sigma^k \Xi + \Lambda) - PT &= \tau^k \left((\sigma^k \Xi + \Lambda)(PT)^{-1} - \frac{I_n}{\tau^k} \right) PT \\ &= \tau^k \left((\sigma^k \Xi + \Lambda)(PT)^{-1} - [\sigma^k\, \sigma_{\max}(\Xi(PT)^{-1}) + \sigma_{\max}(\Lambda(PT)^{-1})] I_n \right) PT \preccurlyeq 0\,, \end{aligned}$$

and (33a) is true. Then, solving (33) for τ^k yields a lower bound for the sequence τ^k, which must satisfy

$$\begin{aligned} [(\alpha - \beta\tau^k)(P^2T + \tau^k P\Upsilon) - \beta(\tau^k)^2(I_n - P)\Upsilon](\tau^{k+1})^2 + (\tau^k)^2(\beta P^2 T + \alpha(I_n - P)\Upsilon)\tau^{k+1} \\ -\alpha(\tau^k)^2 P^2 T \succcurlyeq 0. \end{aligned}$$

A close inspection (omitted in this study) of the above condition yields an upper bound for the sequence (τ^k), which decreases with asymptotic convergence rate $O(2\sigma_{\max}(\Upsilon^{-1}P^2T)/k)$. In view of this conjecture, we can maximize the sequence by balancing the eigenspectrum of the matrix $\Upsilon^{-1}P^2T$. We do so by choosing $T = Q\Upsilon$, where $Q = \mathrm{diag}(q_1 I_{n_1}, \ldots, q_p I_{n_p})$, and the condition becomes

$$\begin{aligned} \big[[(\alpha - \beta\tau^k)(P^2Q + \tau^k P) - \beta(\tau^k)^2(I_n - P)](\tau^{k+1})^2 + (\tau^k)^2(\beta P^2 Q + \alpha(I_n - P))\tau^{k+1} \\ -\alpha(\tau^k)^2 P^2 Q \big] \Upsilon \succcurlyeq 0. \end{aligned} \tag{44}$$

where $\alpha = \sigma^{-1}_{\max}(\Xi(PQ\Upsilon)^{-1})$ and $\beta = \sigma_{\max}(\Lambda(PQ\Upsilon)^{-1})\,\alpha$. The stepsize sequence (6) is obtained by letting $Q = P^{-2}$ (i.e., $T = P^{-2}\Upsilon$) and by recursively taking for τ^{k+1} the positive solution of (44). As we proceed to show next, the sequence (6) is decreasing with asymptotic rate $2/k$.

Fix $T = Q\Upsilon$, where $Q = \mathrm{diag}(q_1 I_{n_1}, \ldots, q_p I_{n_p})$, and ensure that (33a) holds by setting $\sigma^k = \alpha/\tau^k - \beta$, where $\alpha = \sigma^{-1}_{\max}(\Xi(PQ\Upsilon)^{-1})$ and $\beta = \sigma_{\max}(\Lambda(PQ\Upsilon)^{-1})\,\alpha$. Then, Condition (33b) reduces to (44), which rewrites as

$$\begin{aligned} [(\alpha - \beta\tau^k)(q_i + \pi_i\tau^k) + \beta\pi_i(1 - \pi_i)(\tau^k)^2](\tau^{k+1})^2 \\ +(\beta q_i - \alpha\pi_i(1 - \pi_i))(\tau^k)^2\tau^{k+1} - \alpha q_i(\tau^k)^2 \geqslant 0, \quad i = 1, \ldots, p. \end{aligned} \tag{45}$$

The discriminant of (45), given by

$$\begin{aligned}\Delta &= 4\alpha^2\left[1+\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k+\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)^2-\left(\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}\right)\frac{\pi_i^2}{q_i}\right]\frac{(\tau^k)^2}{4}\right]q_i^2(\tau^k)^2\\ &= 4\alpha^2\left[\left(1+\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\frac{\tau^k}{2}\right)^2-\left(\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}\right)\frac{\pi_i^2(\tau^k)^2}{4q_i}\right]q_i^2(\tau^k)^2,\end{aligned} \tag{46}$$

is nonnegative for

$$\tau^k \leqslant \frac{2}{\max\left(0,\left(\frac{\beta}{\alpha}-\frac{\pi_i}{q_i}\right)-\frac{\pi_i}{\sqrt{q_i}}\sqrt{\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}},\frac{\pi_i}{\sqrt{q_i}}\sqrt{\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}}-\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\right)}. \tag{C1}$$

Under (C1), Condition (45) reduces to

$$\tau^{k+1} \geqslant \frac{-(\beta q_i-\alpha\pi_i(1-\pi_i))(\tau^k)^2+\sqrt{\Delta}}{2[(\alpha-\beta\tau^k)(q_i+\pi_i\tau^k)+\beta\pi_i(1-\pi_i)(\tau^k)^2]}, \tag{47}$$

holding for $i=1,\ldots,p$. Now, define $\underline{\delta}$, $\bar{\delta}$ and δ as

$$\underline{\delta}=\min_{i\in\{1,\ldots,p\}}\left\{\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right\},\quad \bar{\delta}=\max_{i\in\{1,\ldots,p\}}\left\{\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right\},\quad \delta=\max(|\underline{\delta}|,|\bar{\delta}|). \tag{48}$$

Let $\gamma>1$, and assume that

$$-\frac{\gamma-1}{\gamma}\leqslant\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2-\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k\leqslant\frac{\gamma-1}{\gamma}, \tag{49}$$

or, equivalently,

$$\begin{aligned}\tau^k &\leqslant \left(\frac{\alpha q_i}{2\beta\pi_i^2}\right)\left[\underline{\delta}+\sqrt{\underline{\delta}^2+4\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)\left(\frac{\gamma-1}{\gamma}\right)}\right],\\ \tau^k &\leqslant \left(\frac{\alpha q_i}{2\beta\pi_i^2}\right)\left[\bar{\delta}-\sqrt{\bar{\delta}^2-4\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)\left(\frac{\gamma-1}{\gamma}\right)}\right]\ \text{if } \bar{\delta}>0.\end{aligned} \tag{C2}$$

Using the identities $\sqrt{1+x}\leqslant 1+(x/2)$, and $(1-x)^{-1}\leqslant 1+x+\gamma x^2$ for $x\leqslant 1-1/\gamma$, we find

$$\begin{aligned}&\frac{1}{\tau^k}\left(\frac{-(\beta q_i-\alpha\pi_i(1-\pi_i))(\tau^k)^2+\sqrt{\Delta}}{2[(\alpha-\beta\tau^k)(q_i+\pi_i\tau^k)+\beta\pi_i(1-\pi_i)(\tau^k)^2]}\right)\\ &\overset{(46)}{\leqslant}\frac{1+(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}-\frac{\pi_i^2}{2q_i})\tau^k+\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)^2-\left(\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}\right)\frac{\pi_i^2}{q_i}\right]\frac{(\tau^k)^2}{8}}{1-\left[(\frac{\beta\pi_i^2}{\alpha q_i})(\tau^k)^2-(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha})\tau^k\right]}\\ &\leqslant\left[1+\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}-\frac{\pi_i^2}{2q_i}\right)\tau^k+\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)^2-\left(\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}\right)\frac{\pi_i^2}{q_i}\right]\frac{(\tau^k)^2}{8}\right]\\ &\qquad\times\left[1+\left[\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2-\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k\right]+\gamma\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k-\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2\right]^2\right]\\ &=1-\left(\frac{\pi_i^2}{2q_i}\right)\tau^k+\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2+\gamma\left[\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2-\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k\right]^2\end{aligned}$$

$$
\begin{aligned}
&+\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)^2-\left(\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}\right)\frac{\pi_i^2}{q_i}\right]\frac{(\tau^k)^2}{8}\\
&+\left[\left[\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2-\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k\right]+\gamma\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k-\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2\right]^2\right]\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}-\frac{\pi_i^2}{2q_i}\right)\tau^k\\
&+\left[\left[\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2-\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k\right]+\gamma\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)\tau^k-\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^2\right]^2\right]\\
&\qquad\qquad\times\left[\left(\frac{\pi_i}{q_i}-\frac{\beta}{\alpha}\right)^2-\left(\frac{\pi_i(2-\pi_i)}{q_i}+\frac{2\beta}{\alpha}\right)\frac{\pi_i^2}{q_i}\right]\frac{(\tau^k)^2}{8}\\
&\overset{(48),(49)}{\leqslant}1-\left(\frac{\pi_i^2}{2q_i}\right)\tau^k+\gamma\left[\frac{17\delta^2}{8}+\left(\frac{\delta}{2}+\frac{\beta}{\alpha}\right)\frac{\pi_i^2}{q_i}+\frac{1+2\delta}{8}\left(\frac{\pi_i^2}{q_i}\right)^2\right](\tau^k)^2+\gamma\left(2\delta+\frac{\pi_i^2}{2q_i}\right)\left(\frac{\beta\pi_i^2}{\alpha q_i}\right)(\tau^k)^3\\
&=1-\left(\frac{\pi_i^2}{2q_i}\right)\tau^k+u(\tau^k),
\end{aligned}
\tag{50}
$$

where $u(\epsilon)=O(\epsilon^2)$. The above result can be balanced by setting $q_i=\pi_i^2$ for $i=1,\dots,p$. In particular, it can be seen from (47) and (50) that $\tau^{k+1}\leqslant\tau^k-(1-\epsilon)(\tau^k)^2/2$ whenever

$$
\tau^k\leqslant\left[\frac{(17\delta^2+6\delta+1)\alpha+8\beta}{8\,(4\delta+1)\,\beta}\right]\left[\sqrt{1+\frac{16\alpha\beta(4\delta+1)\,\epsilon}{[(17\delta^2+6\delta+1)\alpha+8\beta]^2\gamma}}-1\right]. \tag{C3}
$$

It follows that the stepsize sequence (7) satisfies (33) and decreases with complexity $\tau^{k+1}\leqslant\tau^k-(1-\epsilon)(\tau^k)^2/2$ on condition that τ^0 satisfies (C1), (C2) and (C3).

Now we show that the tight stepsize rule (6) vanishes with asymptotic behavior $\tau^{k+1}=\tau^k-(\tau^k)^2/2+O((\tau^k)^3)$. To see this, observe that (45) has two fixed points, 0 and $\frac{\alpha}{\beta}$, where 0 is the unique attracting fixed point. Indeed, the positive solution of (45) with $q_i=\pi_i^2$ is given by (6), which reduces for $\tau^k=\frac{\alpha}{\beta}-\epsilon$ to

$$
\tau^{k+1}=\tau^k-\frac{\beta\pi_i+\alpha}{\beta\pi_i+(1-\pi_i)\alpha}\epsilon+O(\epsilon^2). \tag{51}
$$

Hence, stepsize sequence (6) will run away from α/β and decrease towards 0 provided that $\tau^0<\frac{\alpha}{\beta}$. Moreover, we will have $\sigma^k>0$ for all k in (5).

To find the asymptotic rate of τ^k under $q_i=\pi_i^2$ for $i=1,\dots,p$, we apply $\sqrt{1+x}=1+(x/2)+O(x^2)$ and $(1-x)^{-1}\leqslant 1+x+O(x^2)$ to (6). Proceeding as in (50), we find $\tau^{k+1}=\tau^k-(\tau^k)^2/2+O((\tau^k)^3)$. We infer that $\tau^k-2/k=o(1/k)$, and it follows from (5) that $\sigma^k-(\alpha k/2-\beta)=o(1)$, $S^k-[\alpha k(k+1)/4-(k+1)\beta]=o(k)$.[1]

5. Numerical experiments for an optimal transport problem

We consider an optimal transport problem where a number of customers are willing to buy a service available at various sites $1,\dots,p$. The customers are identical within each of m distinct classes indexed by $1,\dots,m$. Site j is characterized by a cost vector $c_j=(c_{1j},\dots,c_{mj})$, where c_{ij} denotes the cost charged to any customer from class i for buying the service at site j ($i=1,\dots,m$; $j=1,\dots,p$). We use the following notation: for $i=1,\dots,m$ and $j=1,\dots,p$,

- $\mu_i\in\mathbb{R}_{\geqslant 0}$: mass of customers from class i,

[1] If $0<a[1]<2$, the solution of the difference equation $a[k+1]=a[k]-\frac{1}{2}a[k]^2$ is $a[k]=\frac{2}{k}+o(1/k)$ for $k\geqslant 1$.

- $\nu_j \in \mathbb{R}_{\geqslant 0}$: capacity at site j,
- $x_{ij} \in \mathbb{R}_{\geqslant 0}$: mass of customers from class i receiving service at site j,
- $x_j = (x_{1j}, \ldots, x_{mj}) \in \mathbb{R}^m_{\geqslant 0}$: service schedule at site j,
- $x = (x_1, \ldots, x_p) \in \mathbb{R}^{mp}_{\geqslant 0}$: global service schedule,
- $c_j = (c_{1j}, \ldots, c_{mj})$: cost vector associated with site j,
- c_{ij}: cost charged to a customer from class i when receiving service at site j.

One would like to the maximize the overall attractiveness of the service by solving

$$\begin{array}{lll} \underset{x}{\text{minimize}} & \sum_{j=1}^p \left[\langle c_j, x_j\rangle + d_j(x_j)\right] & \\ \text{subject to} & \sum_{j=1}^p x_{ij} = \mu_i & (i = 1, \ldots, m) \\ & \sum_{i=1}^m x_{ij} \leqslant \nu_j & (j = 1, \ldots, p) \\ & x_{ij} \geqslant 0 & (i = 1, \ldots, m,\ j = 1, \ldots, p) \end{array} \tag{OT}$$

where $d_j : \mathbb{R}^m_{\geqslant 0} \mapsto \mathbb{R}_{\geqslant 0}$ is a smooth, convex function such that $d_j(x_j)$ models an additional congestion cost around site j under service schedule x_j.

Problem (OT) rewrites as as an instance of (P) in which $n_j \equiv m$ for $j = 1, \ldots, p$, the equality constraint is specified by $b = \mu = (\mu_1, \ldots, \mu_m)$ and $A = (A_1 \ \cdots \ A_p)$ with $A_j = I_{n_j}$, the feasible sets are given for $j = 1, \ldots, p$ by

$$X_j = \{x_j \in \mathbb{R}^m_{\geqslant 0} : \nu_j - ||x_j||_1 \geqslant 0\}, \tag{52}$$

and the objective functions by

$$h_j(x_j) = \langle c_j, x_j\rangle + d_j(x_j) - \frac{1}{2}\|x_j\|^2_{\Upsilon_j}, \quad g_j(x_j) = \mathbb{I}_{X_j}(x_j) + \frac{1}{2}\|x_j\|^2_{\Upsilon_j}, \tag{53}$$

where h_j is smooth convex and g_j is proper convex lower semi-continuous, and $\mathbb{I}_C$ denotes the indicator function of the set C. The stationarity condition for (OT) implies that every primal-dual solution (x, y) satisfies

$$c_{ij} + \nabla_j d_j(x_j) + y_i > 0 \implies x_{ij} = 0, \tag{54}$$

$$x_{ij} > 0 \implies c_{ij} + \nabla_j d_j(x_j) + \delta_j + y_i = 0, \tag{55}$$

for some $\delta_j \geqslant 0$, where δ_j denotes the dual variable relative to the side inequality constraint $\sum_{i=1}^m x_{ij} - \nu_j \leqslant 0$, and y_i relates to the equality constraint $\sum_{j=1}^p x_{ij} - \mu_i = 0$.

Customer guidance through pricing. The interpretation of these dual variables as prices allows the operator to maximize the service quality by assigning unit prices $p_1, \ldots, p_p$ to the service at each site, under the assumption that all customers are willing to be served at lowest cost, i.e., a customer from class i is going to purchase the service at a site j that minimizes their prospective total cost $c_{ij} + p_j$. Indeed, implementing the price profile

$$p_j = \nabla_j d_j(x_j) + \delta_j - a, \tag{56}$$

where a is an arbitrary constant parameter, yields

$$x_{ij} > 0 \implies c_{ij} + p_j + y_i + a = 0, \tag{57}$$

$$c_{ij} + p_j + y_i + a > 0 \implies x_{ij} = 0 \tag{58}$$

or, equivalently,

$$x_{ij} > 0 \Longrightarrow j \in \arg\min_{j'}\{c_{ij'} + p_{j'}\}. \tag{59}$$

which is the desired result. Any indecisions in (59) may be dealt with in practice using a booking system.

(m,p) :	$(10,10)$	$(20,20)$	$(50,50)$	$(100,100)$	$(10,40)$	$(10,250)$	$(10,1000)$
Constant ($\sigma=1$):	261	519	1 091	1 927	914	3 929	9 350
Constant (f.t. σ):	221	99	44	56	41	96	278
Accelerated:	130	128	152	122	107	92	62

(A) Epochs until $\|Ax-b\|_\infty < 10^{-6}$

(m,p) :	$(10,10)$	$(20,20)$	$(50,50)$	$(100,100)$	$(10,40)$	$(10,250)$	$(10,1000)$
Constant ($\sigma=1$):	409	816	1 988	3 861	1 655	10 762	/
Constant (f.t. σ):	221	99	53	64	45	141	472
Accelerated:	1 589	1 288	2 333	1 094	1 092	1 771	1 773

(B) Epochs until $\mathrm{dist}_\infty\big((\partial_x,-\partial_y)L(x^k,y^k),0\big) < 10^{-6}$ where $L(x,y)=h(x)+g(x)+\langle y,Ax-b\rangle$

TABLE 1. Application of Algorithm 1 to (OT) with $d_j(x_j)=\frac{1}{2}\|x_j\|^2$. Constant (3) with $\sigma=1$ and with finely-tuned σ, and accelerated (6) scaling.

Experiments. In our numerical experiments we use the basic congestion model $d_j(x_j)=\frac{1}{2}\|x_j\|^2_{\Upsilon_j}$ for $j=1,\dots,p$, where $\Upsilon_j \succcurlyeq 0$, which implies convexity in the sense of Assumption 2.1(i) with characteristic $\Upsilon=\mathrm{diag}(\Upsilon_1,\dots,\Upsilon_p)$. In particular, if we set $\Upsilon_j=M_jI_{n_j}$ for some $M_j \geqslant 0$, the primal step on Line 2 of Algorithm 1 reduces to

$$x_j^{k+1} \;=\; \mathrm{prox}_{\mathbb{I}_{X_j}}\Big(\big(M_j+\lambda_j^k\big)^{-1}\big(\lambda_j^k x_j^k-\big(c_j+y^k\big)\big)\Big), \tag{60}$$

where we have used $P_j=(1/\pi_j)\,I_{n_j}$ and either the constant preconditiong $T_j^k=T_j$ with $T_j=(1/\tau_j+\sigma)I_{n_j}$ derived from (3), or the decreasing sequence $T_j^k=T_j/\tau^k$ with $T_j=(M_j/\pi_j^2)I_{n_j}$ and τ^k satisfying (6), which respectively give $\lambda_j^k=(1/\pi_j)(1/\tau_j+\sigma)$ and $\lambda_j^k=(M_j\pi_j/\tau^k)$. The proximal operation in (60) thus involves a (scaled) gradient descent and an Euclidean projection on X_j, which are straightforward to compute.

In our tests, we set $M_{1j}=\cdots=M_{mj}=1$ for $j=1,\dots,p$. All the entries of $c_1,\dots,c_p$, $\mu_1,\dots,\mu_m$, and $\nu_1,\dots,\nu_p$ are picked randomly, independently and uniformly in $[0,1]$. The populations $\mu_1,\dots,\mu_m$ are then normalized so that the total load over total capacity saturation factor is 0.8 i.e., $\sum_{i=1}^m \mu_i = 0.8\sum_{j=1}^p \mu_j$. Algorithm 1 is then implemented with the following block selection policy: at every step, each agent updates their primal variables independently with probability $1/p$. Ignoring all update-free steps, which occur with probability $\pi_0=(1-1/p)^p$, we find $\pi_1=\cdots=\pi_p=[p(1-\pi_0)]^{-1}$ and $\pi_{i,j}=[p^2(1-\pi_0)]^{-1}$ if $i\neq j$. For the constant stepsize policy, we fix σ and we set $\tau_1=\cdots=\tau_p=(1/2\sigma)\,p(1-\pi_0)\,(\sigma_{\max}(\Xi)-1)^{-1}$ in accordance with (3). For the accelerated implementation, the decreasing stepsize sequence τ^k is computed as in (6) with $\tau^0=1$. Since in this problem $\Lambda=0$, we have $\beta=\kappa=0$.

Table 1 displays the number of updates per agent (epochs) that were needed to approach optimality. As stopping criteria we consider the feasiblity residual $\|Ax-b\|_\infty$ and the KKT residual $\mathrm{dist}_\infty\big((\partial_x,-\partial_y)L(x^k,y^k),0\big)$, where $L(x,y)=h(x)+g(x)+\langle y,Ax-b\rangle$ denotes the Lagrangian of (P). In our simulations, the constant scaling implementation proved quite sensitive to the chosen parameters. Setting $\sigma=1$ was fine with smaller networks, though performance deteriorated as the size of the problem increased, and in particular for a larger number p of agents. Speed of convergence can be considerably increased by fine tuning the

parameter values, as can be seen on the second lines of Table 1a and Table 1b, where we used $\sigma = 0.1$ for $(m, p) = (10, 10)$ and $\sigma = 0.01$ for the larger networks. The accelerated scaling, on the other hand, offered good and consistent (near dimension-free) performance even in the larger networks with tens of thousands of variables and many agents.

Appendix A. Proofs of Theorems 2.2 and 2.3

A.1. **Convex case ($\Upsilon = 0$).** If $\Upsilon = 0$, then we take the constant stepsizes $\tau^k = 1$ and $\sigma^k = \sigma$ all k, so that (33a) reduces to a condition on σ,

$$\sigma\Xi + \Lambda \preccurlyeq PT. \tag{61}$$

Besides, in Algorithm 2 we get $S^k = (k+1)\sigma$, while $F_k = \hat{G}^k - G^* + \sigma k(f(s^k) - f^*)$ and $V^k(x) = \frac{\sigma}{2}\|x^k - x\|^2_{P^2T} + k\sigma F_k$.

We are now in a position to show Theorem 2.2. Part (i) in the proof, in particular, follows the lines of (Luke and Malitsky, 2018, proof of Theorem 1, part (i)), with a few changes required by the presence of the smooth component h. All the developments are detailed below for completeness.

Proof of Theorem 2.2. (i) First suppose there is a Lagrange multiplier. Then, (39) becomes

$$kF_k \geqslant \frac{\sigma k^2}{2}(f(s^k) - f^*) - \frac{\|Av^*\|^2}{\sigma} \geqslant -\frac{\|Av^*\|^2}{\sigma}, \tag{62}$$

and we have $V^k(x) \geqslant -\|Av^*\|^2$. The supermartingale theorem applies, and (43) reduces to

$$\sum_{k=0}^{\infty}\left[\frac{1}{2}\|\hat{x}^{k+1} - x^k\|^2_{PT-(\sigma\Xi+\Lambda)} + \sigma(f(x^k) - f^*)\right] < \infty \quad \text{a.s.}, \tag{63}$$

which almost surely yields $f(x^k) - f^* = o(1/k)$ and $\|\hat{x}^{k+1} - x^k\|^2_{PT-(\sigma\Xi+\Lambda)} \to 0$ (hence, by (28), $f(\hat{x}^{k+1}) - f^* = \frac{1}{2}\|\hat{x}^{k+1} - x^*\|^2_{A^\top A} \to 0$), and the sequence $V^k(x^*)$ converges almost surely to a random quantity not smaller than $-\|Av^*\|^2$. It follows that $V^k(x^*)$ is almost surely bounded from above by a constant $C(\omega)$, so that

$$C(\omega)+\|Av^*\|^2 \geqslant V^k(x^*)+\|Av^*\|^2 = \frac{\sigma}{2}\|x^k-x^*\|^2_{P^2T}+\sigma kF_k+\|Av^*\|^2 \overset{(62)}{\geqslant} \frac{\sigma}{2}\|x^k-x^*\|^2_{P^2T} \geqslant 0 \text{ a.s.}, \tag{64}$$

which shows that the sequence x^k is pointwise almost surely bounded, and so are the sequences s^k and z^k as a consequence of (4). Since almost surely $k\sigma F_k = V^k(x^*) - \frac{\sigma}{2}\|x^k - x^*\|^2_{P^2T} \leqslant V^k(x^*)$ is bounded from above, we infer from (62) that $f(s^k) - f^* = O(1/k^2)$ almost surely. Since we also have $f(x^k) - f^* = o(1/k)$ almost surely, $S^{k-1}/S^k = O(1)$, and $\sigma^k/S^k = O(1/k)$ as $k \to \infty$, it follows from Line 1 of Algorithm 2 and from the convexity of f that $f(z^k) - f^* = O(1/k^2)$ almost surely. Then, using $Ax^* = b$, we find

$$\begin{aligned}\langle \nabla f(z^k), x^* - \hat{x}^{k+1}\rangle &\overset{(1)}{=} \langle A^\top(Az^k - b), x^* - \hat{x}^{k+1}\rangle \\ &= \langle A^\top A(z^k - x^*), x^* - \hat{x}^{k+1}\rangle \leqslant \|A(z^k - x^*)\|\|A(x^* - \hat{x}^{k+1})\| \\ &\overset{(28)}{=} 2\sqrt{f(z^k) - f^*}\,\sqrt{f(\hat{x}^{k+1}) - f^*} = o(1/k),\end{aligned} \tag{65}$$

where we have used $f(\hat{x}^{k+1}) - f^* \to 0$. By convexity and smoothness of h, we find

$$\begin{aligned}\langle \nabla h(x^k), x^* - \hat{x}^{k+1}\rangle &= \langle \nabla h(x^k), x^* - x^k\rangle - \langle \nabla h(x^k), \hat{x}^{k+1} - x^k\rangle \\ &\overset{(2)}{\leqslant} [h(x^*) - h(x^k)] - [h(\hat{x}^{k+1}) - h(x^k) - \tfrac{1}{2}\|\hat{x}^{k+1} - x^k\|_\Lambda^2] \\ &= h(x^*) - h(\hat{x}^{k+1}) + \tfrac{1}{2}\|\hat{x}^{k+1} - x^k\|_\Lambda^2.\end{aligned} \tag{66}$$

Consider a bounded sequence (x^k) (this event happens with probability one), and one of its convergent subsequence (x^{k_l}) with accumulation point $\tilde{x}$, which is feasible since $f(\tilde{x}) = \lim_{l\to\infty} f(x^{k_l}) = f^*$. Now, applying Lemma 3.1 at point x^* gives

$$\begin{aligned}g(\hat{x}^{k_l+1}) - g(x^*) &\overset{(14)}{\leqslant} \langle \nabla h(x^{k_l}) + S^{k_l}\nabla f(z^{k_l}) + PT^{k_l}(\hat{x}^{k_l+1} - x^{k_l}), x^* - \hat{x}^{k_l+1}\rangle \\ &\overset{(66)}{\leqslant} h(x^*) - h(\hat{x}^{k_l+1}) + \tfrac{1}{2}\|\hat{x}^{k_l+1} - x^{k_l}\|_\Lambda^2 \\ &\qquad + \langle S^{k_l}\nabla f(z^{k_l}) + PT^{k_l}(\hat{x}^{k_l+1} - x^{k_l}), x^* - \hat{x}^{k_l+1}\rangle,\end{aligned}$$

which rewrites as

$$\begin{aligned}G(\hat{x}^{k_l+1}) - G(x^*) \leqslant \tfrac{1}{2}\|\hat{x}^{k_l+1} - x^{k_l}\|_\Lambda^2 + S^{k_l}\langle \nabla f(z^{k_l}), x^* - \hat{x}^{k_l+1}\rangle \\ + \langle PT^{k_l}(\hat{x}^{k_l+1} - x^{k_l}), x^* - \hat{x}^{k_l+1}\rangle.\end{aligned} \tag{67}$$

Since $\|\hat{x}^{k_l+1} - x^{k_l}\|^2_{PT-(\sigma\Xi+\Lambda)} \to 0$, the sequence $(\hat{x}^{k_l+1})$ is bounded and, by taking the limit of the last equation for $l \to \infty$, we find, using (65), $S^{k_l} = k_l\sigma$ and the lower-semicontinuity of G: $G(\tilde{x}) \leqslant \lim_{l\to\infty} G(\hat{x}^{k_l+1}) \leqslant G(x^*)$, and $x^* \in \mathcal{S}$. To show that the bounded sequence (x^k) has a unique accumulation point, we consider a second convergent subsequence $x^{\bar{k}_l} \to \bar{x}$. Recall that $V^k(x) = \frac{\sigma}{2}\|x^k - x\|^2_{P^2T} + k\sigma F_k$ and that $V^k(x)$ is convergent for our bounded sequence (x^k), thus admitting the same limit for subsequences (x^{k_l}) and $(x^{\bar{k}_l})$. Computing these limits at $x = \tilde{x}$ yields $\lim_{l\to\infty} V^{k_l}(\tilde{x}) = \lim_{l\to\infty} k_l\sigma F_{k_l}$ and $\lim_{l\to\infty} V^{\bar{k}_l}(\tilde{x}) = \frac{\sigma}{2}\|\bar{x} - \tilde{x}\|^2_{P^2T} + \lim_{l\to\infty} \bar{k}_l\sigma F_{\bar{k}_l}$, hence $\lim_{l\to\infty} k_l\sigma F_{k_l} = \frac{\sigma}{2}\|\bar{x} - \tilde{x}\|^2_{P^2T} + \lim_{l\to\infty} \bar{k}_l\sigma F_{\bar{k}_l}$. Repeating this operation at $x = \bar{x}$ gives $\frac{\sigma}{2}\|\tilde{x} - \bar{x}\|^2_{P^2T} + \lim_{l\to\infty} k_l\sigma F_{k_l} = \lim_{l\to\infty} \bar{k}_l\sigma F_{\bar{k}_l}$. Thus $\|\tilde{x} - \bar{x}\|^2_{P^2T} = 0$ and $\tilde{x} = \bar{x}$, which shows that almost surely (x^k) converges pointwise to a point of $\mathcal{S}$. Let x^* be that limit point. From Lemma 3.2, we know that $\sum_{l=0}^k |\Gamma^{k,l}| = \sum_{l=0}^k \Gamma^{k,l} = I_n$. Moreover, (17) gives $\Gamma^{k+1,l} = (S^l/S^k)\Gamma^{l+1,l} = O(1/k)$ and thus, for all l, $\Gamma^{k+1,l} \to 0$ as $k \to \infty$. It follows from (16) and the Toeplitz Theorem, Toeplitz (1911), that $s^k = \sum_{l=0}^k \Gamma^{k,l}x^l \to x^* \in \mathcal{S}$ almost surely, which completes the proof of (i).

(ii) We know from Section 3.7 that almost surely $W^k(x^*)$ converges to $Z(\omega) \geqslant G_{\min} - G^*$ and is thus bounded from above by a constant $C(\omega)$, so that $C(\omega) \geqslant W^k(x^*) \geqslant F_k$. It follows from (41) that $f(s^k) - f^* \leqslant [F_k + G^* - G_{\min}]/(k\sigma) \leqslant [C(\omega) + G^* - G_{\min}]/(k\sigma)$ and, consequently, we find $f(s^k) - f^* = O(1/k)$ almost surely. From (43) we also have $(f(x^k) - f^*) = o(1)$ almost surely. It follows from Line 1 in Algorithm 2 and from the convexity of f that almost surely $f(z^k) - f^* = O(1/k)$.

Now, consider the sublevel sets defined by

$$\mathcal{L}(c) = \{x : \max(f(x) - f^*, G(x) - G^*) \leqslant c\}, \qquad c \geqslant 0. \tag{68}$$

Since $\mathcal{L}(0) = \mathcal{S}$ is bounded and both f and G are convex functions, the sets $\mathcal{L}(c)$ are bounded for all $c \geqslant 0$. From $\hat{G}^k \geqslant G(s^k)$ and the definitions of $W^k(x^*)$, we find $W^k(x) \geqslant G(s^k) - G^* + S^{k-1}(f(s^k) - f^*)$, which yields

$$f(s^k) - f^* \leqslant \frac{1}{S^{k-1}}(W^k(x^*) + G^* - G_{\min}) \quad \text{and} \quad G(s^k) - G^* \leqslant W^k(x^*), \tag{69}$$

in which $W^k(x^*) \to Z(\omega)$ for any outcome $\omega \in \Omega$, where Ω is an event of probability 1. It follows that both sequences $(f(s^k(\omega)) - f^*)$ and $(G(s^k(\omega)) - G^*)$ are bounded by a constant $C'(\omega)$. Hence, the sequence $s^k(\omega)$ belongs to $\mathcal{L}(C'(\omega))$, which is a bounded set. Thus, the sequence $s^k(\omega)$ is bounded if $\omega \in \Omega$.

Since $S^{k-1} = \sigma k$, (69) also yields $f(s^k(\omega)) - f^* = O(1/k)$ if $\omega \in \Omega$. Hence $f(s^k(\omega)) - f^* = O(1/k)$ almost surely, as seen previously. Besides, all the accumulation points of $(s^k(\omega))$ are feasible if $\omega \in \Omega$. It follows that $\liminf_{k\to\infty} G(s^k) \geqslant G^*$ if $\omega \in \Omega$, hence with probability one. Now, observe that (36) reduces in the convex case to

$$\mathbb{E}\left[\frac{1}{2k}\|x^k - x^*\|^2_{P^2T+P\Upsilon}\right] + \sigma k\mathbb{E}[f(s^k) - f^*] + \mathbb{E}[G(s^k) - G^*] \leqslant \frac{C}{\sigma k}, \tag{70}$$

and suppose that there is an event $\tilde{\Omega}$ with nonzero probability such that $\limsup_{k\to\infty} G(s^k(\omega)) > G^*$. Then, taking the limit superior in (70) leads to a contradiction. Hence, $\limsup_{k\to\infty} G(s^k) \leqslant G^*$ with probability one. Thus, $\lim_{k\to\infty} G(s^k) = G^*$ with probability one, and almost surely the accumulation points of (s^k) belong to $\mathcal{S}$.

It remains to derive the convergence rate of the sequence $(f(s^k(\omega)))$. We know that $G(s^k(\omega)) - G^* = o(1)$ for every $\omega \in \Omega$. Consider the nonnegative sequence $u^k(\omega) = \max(0, G^* - G(s^k(\omega)))$. For all k and for all $\omega \in \Omega$, we have $\mathbb{E}[G^* - G(s^k(\omega))|\omega \in \Omega] \leqslant \mathbb{E}[u^k(\omega)|\omega \in \Omega]$ where (u^k) converges pointwise towards 0 on Ω. Since (u^k) satisfies $|u^k(\omega)| \leqslant G^* - G_{\min}$ for all k, with $\mathbb{E}[|G^* - G_{\min}||\omega \in \Omega] = G^* - G_{\min} < \infty$, Lebesgue's dominated convergence theorem applies, and we find $\lim_{k\to\infty} \mathbb{E}[u^k(\omega)|\omega \in \Omega] = 0$. It follows that $\lim_{k\to\infty} \mathbb{E}[G^* - G(s^k(\omega))] \leqslant \text{Prob}(\omega \in \Omega) \lim_{k\to\infty} \mathbb{E}[u^k(\omega)|\omega \in \Omega] + (1 - \text{Prob}(\omega \in \Omega)) \lim_{k\to\infty} \mathbb{E}[u^k(\omega)|\omega \notin \Omega] \leqslant 0$, where we have used $\text{Prob}(\omega \in \Omega) = 1$ and $u^k \leqslant G^* - G_{\min}$. Combining this result with (70) gives

$$\mathbb{E}[f(s^k) - f^*] \leqslant \frac{C}{(\sigma k)^2} + \frac{\mathbb{E}[G^* - G(s^k)]}{\sigma k} = o(1/k),$$

which completes the proof of Theorem 2.2. □

A.2. **Strongly convex case.** The proof of 2.3 under strong convexity is similar to that of Theorem 2.2. Recall from Section 4, however, that in Algorithm 2 we now have $\tau^k - 2/k = o(1/k)$, $\sigma^k - (\alpha k/2 - \beta) = o(1)$, and $S^k - [\alpha k(k+1)/4 - (k+1)\beta] = o(k)$.

Proof of Theorem 2.3. (i) We begin as in the proof of Theorem 2.3(i) by showing that kF_k is bounded from below. If (P') admits an optimal primal-dual pair (x^*, v^*) satisfying (38), then, it follows from (39) that the sequence $V^k(x) \geqslant -\|Av^*\|^2$ is bounded from below, and the supermartingale theorem applies in (34) to the nonnegative sequence $V^k(x) + \|Av^*\|^2$, so that

$$\sum_{k=0}^{\infty}\left[\frac{\sigma^k}{2\tau^k}\|\hat{x}^{k+1} - x^k\|^2_{PT-\tau^k(\sigma^k\Xi+\Lambda)} + (\sigma^k)^2(f(x^k) - f^*)\right] < \infty \quad \text{a.s.}, \tag{71}$$

where $PT - \tau^k(\sigma^k\Xi + \Lambda) \to PT - \alpha\Xi$, which almost surely yields $f(x^k) - f^* = o(1/k^3)$ and $k^2\|\hat{x}^{k+1} - x^k\|^2_{PT} \to 0$ (hence, by (28), $k^2[f(\hat{x}^{k+1}) - f^*] = \frac{k^2}{2}\|\hat{x}^{k+1} - x^*\|^2_{A^\top A} \to 0$), and the sequence $V^k(x^*)$ converges almost surely to a random quantity not smaller than $-\|Av^*\|^2$. It follows that $V^k(x^*)$ is almost surely bounded from above by a constant C, so that

$$C + \|Av^*\|^2 \geqslant V^k(x^*) + \|Av^*\|^2 \overset{(39)}{\geqslant} \frac{\sigma^k}{2\tau^k}\|x^k - x^*\|^2_{P^2T} \geqslant 0 \quad \text{a.s.}, \tag{72}$$

which shows that $\|x^k - x^*\|_{P^2T} = O(1/k)$ and, since P^2T has full rank, the sequence (x^k) converges pointwise almost surely to the solution x^*. Since (17) in Lemma 3.2 gives $\Gamma^{k+1,l} = (S^l/S^k)\Gamma^{l+1,l} = O(1/k^2)$ for all l, so do sequences (s^k) and (z^k) as a consequence of (16), Line 1 in Algorithm 2, and the Toeplitz Theorem. Since almost surely $S^{k-1}F_k = V^k(x^*) - \frac{\sigma^k}{2\tau^k}\|x^k - x^*\|^2_{P^2T+\tau^k P\Upsilon} \leqslant V^k(x^*)$ is bounded from above, we infer from (39) that $f(s^k) - f^* = O(1/k^4)$ almost surely. Since we also have $f(x^k) - f^* = o(1/k^3)$ almost surely, $S^{k-1}/S^k = O(1)$, and $\sigma^k/S^k = O(1/k)$ as $k \to \infty$, it follows from Line 1 of Algorithm 2 and from the convexity of f that $f(z^k) - f^* = O(1/k^4)$ almost surely.

(ii) In the strongly convex case, we know from Section 3.7 that almost surely $W^k(x^*)$ converges to $Z(\omega) \geqslant G_{\min} - G^*$ and is thus bounded from above by a constant $C(\omega)$, so that $C(\omega) \geqslant W^k(x^*) \geqslant F_k$. It follows from (41) that $f(s^k) - f^* \leqslant [F_k + G^* - G_{\min}]/S^{k-1} \leqslant [C(\omega) + G^* - G_{\min}]/S^{k-1}$ and, since $S^k = O(k^2)$, we find $f(s^k) - f^* = O(1/k^2)$ almost surely. From (43) we also have $(f(x^k) - f^*) = o(1/k)$ almost surely. It follows from Line 1 in Algorithm 2 and from the convexity of h that $f(z^k) - f^* = O(1/k^2)$ almost surely.

Since (69) still holds, we find that the sequence $s^k(\omega)$ is bounded for $\omega \in \Omega$, with $f(s^k(\omega)) - f^* = O(1/k^2)$.

All the accumulation points of $(s^k(\omega))$ are thus feasible if $\omega \in \Omega$. As a consequence, we find $\liminf_{k\to\infty} G(s^k) \geqslant G^*$ almost surely. Under strong convexity, (36) now reduces to

$$\mathbb{E}\left[\frac{1}{2S^{k-1}}\|x^k - x^*\|^2_{\Psi^k}\right] + S^{k-1}\mathbb{E}[f(s^k) - f^*] + \mathbb{E}[G(s^k) - G^*] \leqslant \frac{C}{S^{k-1}}. \tag{73}$$

Reasoning as we did for (70), we also find $\limsup_{k\to\infty} G(s^k) \leqslant G^*$ with probability one. It follow that, almost surely, $\lim_{k\to\infty} G(s^k) = G^*$ and x^* is the unique accumulation point of (s^k). As for the convergence rate of the sequence $(f(s^k(\omega)))$, it still holds that $G(s^k(\omega)) - G^* = o(1)$ almost surely, and that $\lim_{k\to\infty} \mathbb{E}[G^* - G(s^k(\omega))] \leqslant 0$. Applying this last resul to (70) yields

$$\mathbb{E}[f(s^k) - f^*] \leqslant \frac{C}{(S^{k-1})^2} + \frac{\mathbb{E}[G^* - G(s^k)]}{S^{k-1}} = o(1/k^2),$$

which completes the proof of Theorem 2.3. □

References

Chambolle, A., Pock, T., 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision 40, 120–145. URL: https://doi.org/10.1007/s10851-010-0251-1, doi:10.1007/s10851-010-0251-1.

Fercoq, O., Richtárik, P., 2015. Accelerated, parallel, and proximal coordinate descent. SIAM Journal on Optimization 25, 1997–2023. URL: https://doi.org/10.1137/130949993, doi:10.1137/130949993, arXiv:https://doi.org/10.1137/130949993.

Luke, D.R., Malitsky, Y., 2018. Block-coordinate primal-dual method for nonsmooth minimization over linear constraints, in: Giselsson, P., Rantzer, A. (Eds.), Large-Scale and Distributed Optimization. Springer International Publishing, Cham, pp. 121–147. URL: https://doi.org/10.1007/978-3-319-97478-1_6, doi:10.1007/978-3-319-97478-1_6.

Malitsky, Y., 2019. The primal-dual hybrid gradient method reduces to a primal method for linearly constrained optimization problems. URL: https://arxiv.org/abs/1706.02602, arXiv:1706.02602.

Richtárik, P., Takáč, M., 2014. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Mathematical Programming 144, 1–38. URL: https://doi.org/10.1007/s10107-012-0614-z.

Richtárik, P., Takáč, M., 2016. Parallel coordinate descent methods for big data optimization. Mathematical Programming 156, 433–484. URL: https://doi.org/10.1007/s10107-015-0901-6.

Toeplitz, O., 1911. Über allgemeine lineare mittelbildungen. Prace Matematyczno-Fizyczne 22, 113–119. URL: http://eudml.org/doc/215310.

Tseng, P., 2008. On accelerated proximal gradient methods for convex-concave optimization. Submitted to SIAM J. Optim.