Contractivity and linear convergence in bilinear saddle-point problems: An operator-theoretic approach

Colin Dirren* ETH Zürich
Mattia Bianchi* ETH Zürich
Panagiotis D. Grontas ETH Zürich
John Lygeros ETH Zürich
Florian Dörfler ETH Zürich

Abstract

We study the convex-concave bilinear saddle-point problem $\min_x \max_y f(x)+y^\top Ax-g(y)$, where both, only one, or none of the functions f and g are strongly convex, and suitable rank conditions on the matrix A hold. The solution of this problem is at the core of many machine learning tasks. By employing tools from monotone operator theory, we systematically prove the contractivity (in turn, the linear convergence) of several first-order primal-dual algorithms, including the Chambolle–Pock method. Our approach results in concise proofs, and it yields new convergence guarantees and tighter bounds compared to known results.

1 INTRODUCTION

We consider the bilinear saddle-point problem

$$\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} \; f(x)+y^\top Ax-g(y), \tag{1}$$

where $f:\mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ and $g:\mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ are proper, convex and closed functions, and $A\in\mathbb{R}^{m\times n}$ is a coupling matrix. The related primal problem is

$$\min_{x\in\mathbb{R}^n} \; f(x)+g^*(Ax), \tag{2}$$

where g^* is the Fenchel conjugate of g. Although many problems can be directly solved using (2), the saddle-point formulation (1) has favorable properties that allow for efficient or parallel solution, e.g., in case of finite-sum (Wang and Xiao, 2017) or sparsity (Lei et al., 2017) structure. For this reason, problem (1) is widely employed, including in empirical risk minimization (Zhang and Lin, 2015; Shalev-Shwartz and Zhang, 2014), unsupervised learning (Xu et al., 2004), reinforcement learning (Du et al., 2017), robust optimization (Ben-Tal et al., 2009), inverse imaging tasks (Chambolle and Pock, 2011), extensive-form games (Farina et al., 2019) and compressed sensing (Bach et al., 2008; Fan et al., 2014).

Here, we aim at solving (1) with linear iteration complexity, under suitable assumptions. More specifically, we are interested in algorithms that are *contractive*. Contractivity is a highly desirable property, that not only implies Q-linear convergence, but also ensures strong robustness properties, with respect to parameters and data, for instance in case of inexact updates (see Section 1.3 below or Chapter 3 of Bullo (2024)). To this goal, we will consider either of the three following alternative conditions:

C1. f is μ_f-strongly convex, g is μ_g-strongly convex, $\mu_f>0$, $\mu_g>0$.

C2. g is μ_g-strongly convex and L_g-smooth, $\mu_g>0$, $L_g>0$; $\mu_A \coloneqq \lambda_{\min}\left(A^\top A\right)>0$.

C3. f is L_f-smooth, g is L_g-smooth, $L_f\geq 0$, $L_g\geq 0$; $n=m$ and $\mu_A \coloneqq \lambda_{\min}\left(A^\top A\right)=\lambda_{\min}\left(AA^\top\right)>0$.

Each of these conditions ensures the strong convexity of the primal problem (2), and also the existence of a unique solution $(x^\star,y^\star)$ to problem (1) (see a proof in Appendix A.1). Furthermore, each condition is tight, in the sense that uniqueness of a primal-dual solution to (1) cannot be guaranteed if any of the subconditions is relaxed—in which case, contractivity is excluded. Conditions C1, C2 and C3 are relevant for a variety of applications, as exemplified next.

Imaging. A common refinement of the Rudin, Osher and Fatemi (ROF) model for image recovery, particularly useful in avoiding the staircasing effect, is the total variation Huber–ROF model (Nikolova and Ng, 2005; Heise et al., 2013). It is given by $\min_{x\in\mathbb{R}^n} \; \ell_\alpha^{\text{hub}}(\Delta x)+\lambda\|x-\hat{x}\|^2$, where $\hat{x}\in\mathbb{R}^n$ is the (vectorized) noisy image, $\lambda>0$ weighs the trade-off

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. *equal contribution.

between fitting and regularization, $\Delta \in \mathbb{R}^{m\times n}$ is the linear difference operator between adjacent pixels, and ℓ_α^{hub} is the Huber regularization with parameter $\alpha > 0$. Its primal-dual reformulation is (Chambolle and Pock, 2011, Eq. 71)

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} \frac{\lambda}{2}\|x-\hat{x}\|^2 - y^\top \Delta x - \frac{\alpha}{2}\|y\|^2 - \iota_P(y), \quad (3)$$

where ι_P is the indicator function of the polar ball. Primal-dual proximal-based methods have been shown to be extremely efficient for the Huber-ROF model, thanks to the existence of closed-form expressions for the involved proximal operators. Clearly, problem (3) satisfies C1.

Affinely constrained optimization. The constrained convex optimization problem

$$\min_{A^\top y\geq b} g(y) = -\max_{A^\top y \geq b} -g(y), \quad (4)$$

arises in a variety of applications, including support vector machines (Schölkopf and Smola, 2002), constrained regression (Monfort et al., 1982), model predictive control (Rawlings et al., 2017). Its saddle-point formulation is obtained by setting $f(x) = -x^\top b + \iota_{\mathbb{R}^m_{\geq 0}}(x)$ in (1) (note that f is non-smooth, as it is constrained on the non-negative orthant). Often, state-of-the-art algorithms solve (1) instead of (4), since the primal-dual formulation allows for efficient distributed computation (Kovalev et al., 2020) and avoids projecting onto $A^\top y \geq b$ at each iteration (Salim et al., 2022a). Whenever g is smooth and strongly convex and $A^\top$ is full-row rank, C2 is satisfied. This also holds for the case of equality constraints $A^\top y = b$ (e.g., in network flow problems (Jiang and Schizas, 2021)).

Reinforcement Learning. Policy evaluation is a key task in reinforcement learning. It consists in approximating the value function V^π, namely the expected cumulative reward achieved by a given policy π,

$$V^\pi(s) = \mathbb{E}\Big[\sum\nolimits_{t=0}^{\infty}\gamma^t R(s_t,a_t)|s_0 = s, a_t \sim \pi(s_t)\Big],$$

where $\gamma \in (0,1)$ is a discount factor and $R(s_t,a_t)$ is the reward for taking action a_t in state s_t. A common approach for policy evaluation is to use a linear function approximation $V^\pi(s) \approx \phi(s)^\top x$, where ϕ is a feature map of the state and the weights $x \in \mathbb{R}^n$ can be estimated by minimizing the empirical mean squared projected Bellman error, for a dataset of length T (Du et al., 2017):

$$\min_x \ \|Ax-b\|^2_{C^{-1}},$$

with $C = \sum_{t=0}^T \phi(s_t)\phi(s_t)^\top$, $b = \sum_{t=0}^T r_t\phi(s_t)$, $A = C - \gamma\sum_{t=0}^T \phi(s_t)\phi(s_{t+1})^\top$. Using gradient methods to solve this problem would require inverting C. The corresponding primal-dual formulation

$$\min_x \max_y \ -y^\top Ax - \frac{1}{2}\|y\|^2_C - y^\top b, \quad (5)$$

is often preferred, as computing gradients does not require matrix inversion and thanks to its finite-sum structure (Du et al., 2017). Problem (5) is in the form (1), and it satisfies C3 if A is invertible.

1.1 Related work

A large body of research focuses on showing linear convergence of first-order primal-dual algorithms for solving (1), under various assumptions (Moslem Zamani and de Klerk, 2024; Jiang et al., 2022; Zhang et al., 2022; Jiang et al., 2023b), including C1, C2 and C3, see Table 1. For instance, Condition C1 is considered in their seminal paper by Chambolle and Pock (2011) and in a number of extensions (Lorenz and Schneppe, 2023). The Chambolle–Pock algorithm is also studied under C2 by Wang and Xiao (2017). Most literature focuses instead on methods that do not rely on proximal mappings, where both f and g are required to be smooth (Korpelevich, 1976; Du and Hu, 2019): Du and Hu (2019) first prove linear convergence of the primal-dual gradient method under C2 and smoothness of f; C3 is first exploited for the accelerated method of Kovalev et al. (2022). In essence, most existing analysis methods are customized to particular problems. For example, a typical strategy to prove linear convergence is to find a merit function that decreases linearly along the algorithm iterates (Chambolle and Pock, 2011; Wang and Xiao, 2017; Kovalev et al., 2022); yet, finding such a function is not easy, as the choice heavily depends on problem assumptions and algorithm. This approach makes a systematic treatment difficult and provides little insight for the design of new methods. Further, few works focus on contractivity properties, and those that do are limited to Condition C1 or to the special case of primal-dual methods for constrained optimization. For the general case, we are not aware of any contractivity result for problem (1) under C2 or C3.

1.2 Contributions

In this paper, we take a fresh look at linear convergence of first-order primal-dual algorithms through the lens of operator theory. We start by casting several algorithms as preconditioned forward-backward methods. This unified perspective provides a structured way to prove contractivity, and hence linear convergence to the solution of problem (1). Our contributions are the following:

- By leveraging tools from monotone operator theory,

Table 1: The alternative assumptions for linear convergence C1, C2, C3, in the literature.

Asm.	Extra asm.	References	Convergence type
C1		Chambolle and Pock (2011, 2016b); Mokhtari et al. (2020); Jiang et al. (2023a)	R-linear
		Bredies et al. (2022); Balamurugan and Bach (2016); O'Connor and Vandenberghe (2020)	contractivity
	f, g smooth	Arrow et al. (1958);Wang and Li (2020); Kovalev et al. (2022); Korpelevich (1976); Thekumparampil et al. (2022)	Q/R-linear
	f, g smooth	Chen and Rockafellar (1997); Bauschke and Combettes (2017)	contractivity
		this paper	contractivity
C2		Wang and Xiao (2017); Sadiev et al. (2022)	Q/R-linear
	f smooth	Du and Hu (2019); Kovalev et al. (2022); Zhang et al. (2022); Qureshi and Khan (2023)	R-linear
	$f = 0$ (optimization)	Qu and Li (2019); Alghunaim and Sayed (2020); Salim et al. (2022b)	Q-linear
	$f = 0$ (optimization)	Cisneros-Velarde et al. (2022); Su et al. (2019)	contractivity
		this paper	contractivity
C3		Kovalev et al. (2022)	R-linear
	f, g affine or zero	Daskalakis et al. (2018); Mokhtari et al. (2020); Liang and Stokes (2019); Gidel et al. (2019)	Q/R-linear
		this paper	contractivity

we propose a systematic and interpretable approach to derive and establish contractivity of first-order primal-dual algorithms, both proximal and gradient-based, and under any of the conditions C1, C2 or C3.

- In our analysis, we put forward the notion of inverse Lipschitz operator, which to the best of our knowledge has never been considered in saddle-point problems. This allows us, among others, to prove the first linear rate for the Chambolle–Pock algorithm under C3.

- We improve on existing convergence results, by showing both stronger notions of linear convergence (i.e., contractivity rather than Q- or R-linear) and sharper rates. For instance, we improve on the known rate for the (vanilla) Chambolle–Pock algorithm under C1, and we provide the first contractivity results under C2 or C3.

The remainder of this paper is organized as follows. In Section 1.3 we introduce some necessary notations and operator theory basics. In Section 2 we present our operator-theoretic approach. Section 3 contains our main contractivity results for primal-dual algorithms under C1, C2, and C3. We close the paper in Section 4 with some extensions and remarks for future work.

1.3 Preliminaries

Notation. For a symmetric and positive semidefinite matrix $\Phi \in \mathbb{R}^{q\times q} \succcurlyeq 0$, $\langle \omega, u\rangle_\Phi = u^\top \Phi \omega$ and $\|u\|_\Phi^2 = \langle u, u\rangle_\Phi$; we omit the subscripts if $\Phi = I$. The minimum and maximum eigenvalues of a symmetric matrix C are denoted by $\lambda_{\min}(C)$ and $\lambda_{\max}(C)$. We use the convention $\frac{1}{0} = +\infty$. A scalar non-negative sequence $(v^k)_{k\in\mathbb{N}}$ converges Q-linearly (to zero) if $v^{k+1} \leq \rho v^k$ for some $0 < \rho < 1$ and all $k \in \mathbb{N}$; R-linearly if $v^k \leq M\rho^k$ for some $0 < \rho < 1$, $M > 0$, and for all $k \in \mathbb{N}$.

Contractivity. We recall the concept of contractivity, which is central to our work.

Definition 1 (Contractivity)**.** *Let $\Phi \succ 0$. The operator $\mathcal{A} : \mathbb{R}^q \to \mathbb{R}^q$ is contractive in $\|\cdot\|_\Phi$ with parameter (or rate) $0 \leq \rho < 1$ if, for all $\omega, \omega' \in \mathbb{R}^q$,*

$$\|\mathcal{A}(\omega) - \mathcal{A}(\omega')\|_\Phi \leq \rho \|\omega - \omega'\|_\Phi, \tag{6}$$

namely, if $\mathcal{A}$ is ρ-Lipschitz in $\|\cdot\|_\Phi$. We also say that the iteration $\omega^{k+1} = \mathcal{A}(\omega^k)$ is contractive if $\mathcal{A}$ is.

A contractive iteration has a unique fixed point $\omega^\star = \mathcal{A}(\omega^\star)$ (Banach, 1922, Th. 6). Furthermore, it is easily seen that contractivity implies Q-linear (hence, R-linear) convergence with rate ρ of the sequence $(\|\omega^k - \omega^\star\|_\Phi)_{k\in\mathbb{N}}$, but not vice versa. Compared to Q-linear convergent iterations, contractive algorithms enjoy superior numerical stability and robustness (e.g., in case of perturbed updates), tracking properties (e.g., for problems with streaming data), and modularity (the composition of contractive operators is contractive). We illustrate in details the critical distinction between contractivity and Q-linear convergence in Appendix A.2.

Operator theory. (Bauschke and Combettes, 2017) A (set-valued) operator $\mathcal{A} : \mathbb{R}^q \rightrightarrows \mathbb{R}^q$ is characterized by its graph $\mathrm{gra}(\mathcal{A}) := \{(\omega, u) \mid u \in \mathcal{A}(\omega)\}$. The inverse operator $\mathcal{A}^{-1} : \mathbb{R}^q \rightrightarrows \mathbb{R}^q$ is always well-defined via

$\mathrm{gra}(\mathcal{A}^{-1}) = \{(u,\omega) \mid (\omega,u) \in \mathrm{gra}(\mathcal{A})\}$. The operator $\mathcal{A}$ is μ-strongly monotone if there is $\mu > 0$ such that, for all $(\omega,u),(\omega',u') \in \mathrm{gra}(\mathcal{A})$,

$$\langle u-u', \omega-\omega'\rangle \geq \mu\|\omega-\omega'\|^2, \tag{7}$$

and monotone if this holds with $\mu = 0$. Further, $\mathcal{A}$ is maximally monotone if it is monotone and there is no other monotone operator $\bar{\mathcal{A}}$ such that $\mathrm{gra}(\mathcal{A}) \subset \mathrm{gra}(\bar{\mathcal{A}})$. The identity operator is $\mathrm{Id} : \omega \mapsto \omega$. For an extended-value function $\psi : \mathbb{R}^q \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, $\psi^*(y) = \sup\{\langle y,x\rangle - \psi(x) : x \in \mathbb{R}^q\}$ is its Fenchel conjugate; and $\partial\psi : \mathrm{dom}(\psi) \rightrightarrows \mathbb{R}^q$ is its convex subdifferential,

$$\partial\psi(x) = \{v \in \mathbb{R}^q \mid \psi(z) \geq \psi(x) + \langle v, z-x\rangle, \forall z\}. \tag{8}$$

Let ψ be proper, closed, convex. Then $\partial\psi$ is maximally monotone. Furthermore, ψ is L-smooth if and only if $\partial\psi$ is L-Lipschitz continuous (in which case, $\partial\psi$ is single-valued and $\partial\psi = \nabla\psi$) and ψ is μ-strongly-convex if and only if $\partial\psi$ is μ-strongly monotone. The proximal operator of ψ is $\mathrm{prox}_\psi := (\mathrm{Id} + \partial\psi)^{-1}$, and

$$\mathrm{prox}_\psi(\omega') = \mathrm{argmin}_{\omega\in\mathbb{R}^q}\{\psi(\omega) + \tfrac{1}{2}\|\omega'-\omega\|^2\}. \tag{9}$$

Definition 2 (Inverse Lipschitz). *The operator $\mathcal{A} : \mathbb{R}^q \rightrightarrows \mathbb{R}^q$ is $\frac{1}{L}$-inverse Lipschitz if $\|u-u'\| \geq \frac{1}{L}\|\omega-\omega'\|$ for all $(\omega,u),(\omega',u') \in \mathrm{gra}(\mathcal{A})$, namely if $\mathcal{A}^{-1}$ is L-Lipschitz.*

The inverse Lipschitz property was recently used by Ryu et al. (2022) and Gadjov and Pavel (2023); earlier, a local version was introduced by Rockafellar (1976), which is related to many calmness conditions in the literature (e.g., see Jiang et al. (2022)). An inverse Lipschitz operator has at most one zero, and a strongly monotone operator is inverse Lipschitz (as shown in Appendix A.4).

2 UNIFIED ANALYSIS STRATEGY

Algorithm derivation. By the first order optimality conditions, problem (1) is equivalent to finding a zero of the *saddle-point operator* $F : \mathbb{R}^{n+m} \rightrightarrows \mathbb{R}^{n+m}$,

$$F(\omega) = F(x,y) := \begin{bmatrix} \partial f(x) + A^\top y \\ \partial g(y) - Ax \end{bmatrix}, \tag{10}$$

i.e., finding a vector $\omega^\star = (x^\star, y^\star)$ such that $\mathbf{0} \in F(\omega^\star)$. In turn, one way to approach this problem is to employ a preconditioned *forward-backward splitting* (Bauschke and Combettes, 2017, Th. 26.14), namely the iteration

$$\omega^{k+1} = (\mathrm{Id} + \Phi^{-1}F_\mathrm{b})^{-1} \circ (\mathrm{Id} - \Phi^{-1}F_\mathrm{f})(\omega^k) \tag{11}$$

where F is split as $F = F_\mathrm{f} + F_\mathrm{b}$, F_f must be single-valued, $\Phi \in \mathbb{R}^{(n+m)\times(n+m)}$ is a positive definite symmetric *preconditioning* matrix to be designed, and

$$\mathcal{F} := (\mathrm{Id} - \Phi^{-1}F_\mathrm{f}), \quad \mathcal{B} := (\mathrm{Id} + \Phi^{-1}F_\mathrm{b})^{-1}, \tag{12}$$

are the forward step and implicit[1] backward step, respectively. The fixed points of the iteration (11) are the vectors $\omega^\star$ that solve $\mathbf{0} \in \Phi^{-1}F(\omega^\star)$, equivalently $\mathbf{0} \in F(\omega^\star)$, equivalently the solutions of (1). For instance, if f and g are differentiable and we choose $F_\mathrm{f} = F$ (hence, $F_\mathrm{b} = \mathbf{0}$ is the zero operator) and $\Phi = \mathrm{diag}\left(\frac{1}{\tau}I_n, \frac{1}{\sigma}I_m\right)$ for some $\tau, \sigma > 0$, we obtain the iteration

$$x^{k+1} = x^k - \tau(\nabla f(x^k) + A^\top y^k) \tag{13a}$$
$$y^{k+1} = y^k - \sigma(\nabla g(y^k) - Ax^k), \tag{13b}$$

namely the primal-dual gradient (PDG) method. This algorithm is not guaranteed to converge without strong convexity assumptions (for instance, it does not converge if $f = g = 0$ and $A = 1$, which satisfies C3). To derive iterations that converge without extra assumptions and converge linearly if C1, C2, or C3 hold, we instead place the skew symmetric linear operator $(A^\top y, -Ax)$ in the backward step. To obtain inexpensive updates and facilitate the computation of $\mathcal{B}$, we choose the preconditioning matrix as

$$\Phi_{\tau,\sigma} = \begin{bmatrix} \frac{1}{\tau}I_n & -A^\top \\ -A & \frac{1}{\sigma}I_m \end{bmatrix}, \tag{14}$$

where we highlight the dependence on the step sizes τ and σ via the subscript. Note that $\Phi_{\tau,\sigma} \succ 0$ if $\tau\sigma\|A\|^2 < 1$.

For different splittings, i.e., choices of F_b and F_f, we obtain three different algorithms, whose complete derivation we defer to Appendix B. The first algorithm is the Chambolle–Pock method

$$x^{k+1} = \mathrm{prox}_{\tau f}\left(x^k - \tau A^\top y^k\right) \tag{15a}$$
$$y^{k+1} = \mathrm{prox}_{\sigma g}\left(y^k + \sigma A(2x^{k+1} - x^k)\right), \tag{15b}$$

which requires that the proximal operators of both f and g are available in closed form. The second is the semi-implicit method

$$x^{k+1} = \mathrm{prox}_{\tau f}\left(x^k - \tau A^\top y^k\right) \tag{16a}$$
$$y^{k+1} = y^k - \sigma\left(\nabla g(y^k) - A(2x^{k+1} - x^k)\right), \tag{16b}$$

which requires g to be smooth, but only f to be prox-friendly. The third is the fully explicit method

$$x^{k+1} = x^k - \tau\left(\nabla f(x^k) + \tau A^\top y^k\right) \tag{17a}$$
$$y^{k+1} = y^k - \sigma\left(\nabla g(y^k) - A(2x^{k+1} - x^k)\right). \tag{17b}$$

where both f and g are only required to be smooth.

[1] Computing $v = (\mathrm{Id} - \Phi^{-1}F_\mathrm{b})^{-1}(u)$ means solving for v the regularized inclusion $\mathbf{0} \in \Phi(v-u) + F_\mathrm{b}(v)$. Note that $\mathcal{B}$ is single-valued in all our results, see Appendix B.

Main idea. The derivation technique just sketched is itself not novel (Condat, 2013; Vũ, 2013; Combettes et al., 2014; Condat et al., 2023). For instance, it has been known since He and Yuan (2012) that the Chambolle–Pock algorithm can be interpreted as a preconditioned forward-backward method. Instead, our contribution is a novel approach to exploit the operator-theoretic formulation (11) for linear convergence arguments. In particular, our strategy is to prove contractivity properties for the forward and backward steps. Concerning the latter, we will use the notion of inverse Lipschitz operators to deal with the lack of strong convexity, as exemplified in the following proposition, proven in Appendix C.1.

Proposition 1. *Let $F:\mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximally monotone and $\frac{1}{L}$ inverse Lipschitz. Then the resolvent operator $\mathrm{J}_F := (\mathrm{Id}+F)^{-1}$ is $\frac{L}{\sqrt{L^2+1}}$-contractive.*

In fact, one of the major technical novelties of this paper is to show that C2 and C3 induce certain useful inverse Lipschitz properties (see Lemmas 1 and 2 below). Nonetheless, Proposition 1 is still insufficient for the problem at hand. First, the preconditioning introduces technical complications. Second, often in our setup inverse Lipschitz and strong convexity properties do not hold on the whole space (x, y), but only with respect to a subset of the variables; hence, forward and backward steps are not independently contractive. To address both challenges, our enabling result is the following key proposition, which leverages the interplay between the forward and backward steps and exploits "partial contractivity" properties in weighted norms. The proof is found in Appendix C.2.

Proposition 2. *Let $\Phi_{\tau,\sigma} \succ 0$ as in (14) and recall the forward step $\mathcal{F}$ and backward step $\mathcal{B}$ in (12). Assume that there exist a scalar $0 < \gamma \leq 1$ and matrices $\Psi_{\mathrm{b}} \succcurlyeq 0$, $\Psi_{\mathrm{f}} \succcurlyeq 0$, such that, for all $\omega, \omega' \in \mathbb{R}^{n+m}$,*

A1. $\|\mathcal{B}(\omega) - \mathcal{B}(\omega')\|^2_{\Phi_{\tau,\sigma}+\Psi_{\mathrm{b}}} \leq \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}}$

A2. $\|\mathcal{F}(\omega) - \mathcal{F}(\omega')\|^2_{\Phi_{\tau,\sigma}} \leq \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}-\Psi_{\mathrm{f}}}$

and $\Psi_{\mathrm{b}} + \Psi_{\mathrm{f}} \succcurlyeq \gamma(\Phi_{\tau,\sigma} + \Psi_{\mathrm{b}})$. Then, the forward-backward iteration in (11) is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}+\Psi_{\mathrm{b}}}$, with rate $\rho = \sqrt{1-\gamma}$.

By directly exploiting the assumptions on the operators F_{f} and F_{b} (namely, conditions C1, C2, C3), we can guarantee the partial contractivity conditions in A1 and A2 for each of the methods (15), (16), (17). In turn, Proposition 2 allows us to conclude the contractivity of our algorithms, as detailed in Section 3, and informally summarized next, for readability.

Theorem 1. *Under either C1, C2, or C3, there exist stepsizes $\tau, \sigma > 0$ such that:*

- *algorithm* (15) *is contractive;*
- *if g is smooth, algorithm* (16) *is contractive;*
- *if f and g are smooth, algorithm* (17) *is contractive.*

The proofs of our results can be found in Appendix C.

3 MAIN RESULTS

3.1 Chambolle–Pock algorithm

The first method we consider is the Chambolle–Pock algorithm (Chambolle and Pock, 2011) in (15), which is widely used in inverse imaging tasks. The method is recalled in Algorithm 1 below, where we highlight the hyperparameter θ for the sake of later comparison with related work, though we always set $\theta = 1$ here.

Algorithm 1 Chambolle–Pock method

Require: $x^0 \in \mathbb{R}^n, y^0 \in \mathbb{R}^m$, step sizes $\tau, \sigma > 0$, $\theta = 1$
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $\quad x^{k+1} = \mathrm{prox}_{\tau f}\left(x^k - \tau A^\top y^k\right)$
3: $\quad y^{k+1} = \mathrm{prox}_{\sigma g}\left(y^k + \sigma A\left((1+\theta)x^{k+1} - \theta x^k\right)\right)$
4: **end for**

Algorithm 1 can be cast as the forward-backward iteration in (11) by choosing $F_{\mathrm{b}} = F$, $F_{\mathrm{f}} = \mathbf{0}$ and $\Phi_{\tau,\sigma}$ as in (14), see Appendix B.1. It is known that contractivity of Algorithm 1 can be inferred by strong monotonicity of F, when C1 holds (Bredies et al., 2022). Here we provide a refined analysis, that results in improved rates. Further, for analysis under C2 and C3, we provide the following inverse Lipschitz properties, whose derivation is completely novel. The expressions for the constants R_2 and R_3 can be found in the appendix.

Lemma 1. *If C2 holds, the operator F in* (10) *is $\frac{1}{R_2}$-inverse Lipschitz, $R_2 > 0$.*

Lemma 2. *If C3 holds, the operator F in* (10) *is $\frac{1}{R_3}$-inverse Lipschitz, $R_3 > 0$.*

Based on Lemmas 1 and 2 we can guarantee conditions A1 and A2 in Proposition 2, with $\Psi_{\mathrm{f}} = \mathbf{0}$, $\gamma > 0$. From now on, let us use the shorthand

$$\zeta_{\tau,\sigma} := \max\left\{\tfrac{1}{\tau}, \tfrac{1}{\sigma}\right\} + \|A\|. \tag{18}$$

Theorem 2 (Contractivity of Algorithm 1)**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper, convex, closed functions. Let $\phi_{\tau,\sigma}$ be as in* (14)*. Then, the fixed points of Algorithm 1 coincide with the solutions of* (1)*. Furthermore, if $\tau, \sigma > 0$ are chosen such that $\tau\sigma\|A\|^2(1+\epsilon)^2 \leq 1$, for some $\epsilon > 0$, then:*

(i) If C1 holds, then Algorithm 1 is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with rate

$$\rho = \left(1 + \min\{\mu_f\tau, \mu_g\sigma, \kappa\}\right)^{-1} < 1,$$

where

$$\kappa := \frac{\mu_f\tau+\mu_g\sigma-\sqrt{(\mu_f\tau-\mu_g\sigma)^2+4\|A\|^2\mu_f\mu_g\tau^2\sigma^2}}{2(1-\sigma\tau\|A\|^2)} > 0.$$

The best rate is obtained with $\tau = \frac{1}{(1+\epsilon)\|A\|}\sqrt{\frac{\mu_g}{\mu_f}}$ *and* $\sigma = \frac{1}{(1+\epsilon)\|A\|}\sqrt{\frac{\mu_f}{\mu_g}}$*;*

(ii) If C2 holds, then Algorithm 1 is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ *with rate*

$$\rho = \frac{R_2\zeta_{\tau,\sigma}}{\sqrt{(R_2\zeta_{\tau,\sigma})^2+1}} < 1.$$

The best rate is obtained with $\tau=\sigma=\frac{1}{(1+\epsilon)\|A\|}$*;*

(iii) If C3 holds, then Algorithm 1 is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ *with rate*

$$\rho = \frac{R_3\zeta_{\tau,\sigma}}{\sqrt{(R_3\zeta_{\tau,\sigma})^2+1}} < 1.$$

The best rate is obtained with $\tau=\sigma=\frac{1}{(1+\epsilon)\|A\|}$*.*

Let us emphasize that, by definition of contractivity, under C1, C2, or C3, the previous theorem immediately implies Q-linear convergence, i.e.,

$$\|\omega^k-\omega^\star\|_{\Phi_{\tau,\sigma}} \le \rho^k\|\omega^0-\omega^\star\|_{\Phi_{\tau,\sigma}}, \tag{19}$$

for the sequence $(\omega^k)_{k\in\mathbb{N}} = (x^k,y^k)_{k\in\mathbb{N}}$ generated by Algorithm 1, where $\omega^\star$ is the unique solution to (10).

We note that, under C1, the work (Chambolle and Pock, 2011) provides the R-linear convergence rate $\tilde\rho = \left(1+\frac{\sqrt{\mu_f\mu_g}}{\|A\|}\right)^{-\frac{1}{2}}$. Theorem 2 improves on this result, by guaranteeing both Q-linear convergence and a tighter bound. In fact, the optimal step sizes choice in Theorem 2(i) results in the rate $\rho = \left(1+\frac{\sqrt{\mu_f\mu_g}}{(\epsilon+2)\|A\|}\right)^{-1} = \left(1+\frac{2\sqrt{\mu_f\mu_g}}{(\epsilon+2)\|A\|}+\frac{\mu_f\mu_g}{(\epsilon+2)^2\|A\|^2}\right)^{-\frac{1}{2}}$ and, for any μ_f, μ_g, $\|A\|$, we can always choose ϵ small enough to ensure $\rho < \tilde\rho$. Furthermore, when $A=\mathbf{0}$, the step sizes can be chosen as large as desired, and $\rho = (1+\min\{\mu_g\tau,\mu_g\sigma\})^{-1}$, which is the well-known bound for the proximal-point algorithm (see Appendix C.1). Let us note that a different rate was provided in (Chambolle and Pock, 2016b), but for a different choice of $\theta<1$. Also for $\theta<1$, R-linear convergence of Algorithm 1 was proven by Wang and Xiao (2017) under C2. In contrast, Theorem 2(ii) shows contractivity with $\theta=1$. The choice $\theta=1$ is particularly relevant, as in this case the algorithm can be interpreted as a proximal-point method with symmetric preconditioning (He and Yuan, 2012), allowing one to leverage the extensive related theory; for instance, immediately enabling provably convergent accelerations (Bianchi et al., 2022; Iutzeler and Hendrickx, 2019). Finally, Theorem 2(iii) is the first linear convergence rate for the Chambolle–Pock algorithm under C3, namely if f and g are merely convex, but A is invertible. To the best of our knowledge, only Kovalev et al. (2022) studied linear convergence under C3, but based on a gradient-based method, rather than proximal-based.

Remark 1 (Degenerate preconditioner)**.** *The rates in Theorem 2 are decreasing in* ϵ*, thus the best rates are obtained for* $\epsilon\to 0$*. This should be interpreted as an asymptotic result, as for* $\epsilon=0$*, the matrix* $\Phi_{\tau,\sigma}$ *might be singular. In terms of non-weighted norm, one can only conclude the estimate* $\|\omega^k-\omega^\star\| \le \left(C\frac{1}{\epsilon}\right)\rho^k\|\omega^0-\omega^\star\|$*, for some* $C>0$*, which is vacuous if* $\epsilon=0$*. Let us note that the convergence of Algorithm 1 was also shown for degenerate (i.e., singular) preconditioners, with* $\tau\sigma\|A\|^2=1$ *(Condat, 2013), with known linear rates (Chambolle and Pock, 2011), and more recently even for* $\tau\sigma\|A\|^2<4/3$ *(Banert et al., 2023). Extending our results to these cases is an open problem.*

3.2 Semi-implicit primal-dual method

Next, we consider the method in (16), recalled in Algorithm 2, which requires g to be smooth. This method is an instance of the Condat–Vũ splitting (Condat, 2013; Vũ, 2013), applicable to linear regression with ℓ_1 regularization, inverse imaging models (Chambolle and Pock, 2016a), and affine-constrained optimization (Zhu et al., 2023).

Algorithm 2 Semi-implicit primal-dual method

Require: $x^0\in\mathbb{R}^n, y^0\in\mathbb{R}^m$, step sizes $\tau,\sigma>0$, $\theta=1$

1: **for** $k=0,1,\cdots$ **do**
2: $\quad x^{k+1} = \operatorname{prox}_{\tau f}\left(x^k-\tau A^T y^k\right)$
3: $\quad y^{k+1} = y^k-\sigma\left(\nabla g(y^k)-A\left((1+\theta)x^{k+1}-\theta x^k\right)\right)$
4: **end for**

Algorithm 2 corresponds to the iteration in (11), with $F_{\mathrm{b}}(\omega) = (\partial f(x)+A^\top y, -Ax)$, $F_{\mathrm{f}}(\omega) = (\mathbf{0},\nabla g(y))$ and $\Phi_{\tau,\sigma}$ as in (14), see Appendix B.2. Under C3, we can show contractivity of this algorithm by exploiting that F_{b} is inverse Lipschitz, analogously to Theorem 2(iii). Proving linear convergence for Algorithm 2 under C1 and C2 is more complex, as neither $\mathcal{F}$ nor $\mathcal{B}$ are contractions (note further that also the sum $F=F_{\mathrm{f}}+F_{\mathrm{b}}$ is strongly monotone, cf. Theorem 6.1 in Giselsson and Moursi (2021)). Instead, we exploit the intuition that the forward operator F_{f} is strongly monotone, but only with respect to the dual variable y; and the backward operator F_{b} is inverse Lipschitz with respect to the primal variable x. In turn, we can prove properties A1, A2 in Proposition 2, in the following lemmas.

Lemma 3. *Let either C1 or C2 hold (so either* $\mu_f>0$ *or* $\mu_A>0$*), and let* $\tau\sigma\|A\|^2<1$*,* $F_{\mathrm{b}}(\omega) = (\partial f(x)+$

$A^\top y, -Ax)$. Then, A1 *in Proposition 2 holds with $\Psi_{\mathrm{b}} = \mathrm{diag}(\gamma_x I_n, \mathbf{0}_{m\times m})$, where $\gamma_x = \frac{\mu_A}{\zeta_{\tau,\sigma}} + 2\mu_f > 0$.*

Lemma 4. *Let either C1 or C2 hold, and let $\sigma < \frac{2}{2\tau\|A\|^2 + L_g}$, $F_{\mathrm{f}}(\omega) = (\mathbf{0}, \nabla g(y))$. Let $\xi_1 = \frac{1}{\sigma} - \tau\|A\|^2 > 0$, $\xi_2 = \frac{1}{\sigma} - \tau\mu_A > 0$. Then,* A2 *in Proposition 2 holds with $\Psi_{\mathrm{f}} = \mathrm{diag}(\mathbf{0}_{n\times n}, \gamma_y I_m)$, where*

$$0 < \gamma_y = \begin{cases} \frac{2L_g\mu_g}{L_g+\mu_g} + \frac{\mu_g^2(2\xi_1 - L_g - \mu_g)}{(L_g+\mu_g)\xi_2} & \text{if } \xi_1 \geq \frac{L_g+\mu_g}{2} \\ \frac{2L_g\mu_g}{L_g+\mu_g} - \frac{L_g^2(L_g+\mu_g-2\xi_1)}{(L_g+\mu_g)\xi_1} & \text{if } \xi_1 < \frac{L_g+\mu_g}{2}. \end{cases}$$

When $A = \mathbf{0}$, Lemma 4 simplifies to

$$\begin{aligned} &\|y - \sigma\nabla g(y) - (y - \sigma\nabla g(y'))\|^2 \\ &\leq \max\left\{|1-\sigma\mu_g|^2, |1-\sigma L_g|^2\right\}\|y - y'\|^2 \end{aligned}$$

for all $\sigma < \frac{2}{L_g}$, which is the usual contractivity result for gradient descent (see Proposition 5 in Appendix C.1). The following is the main result of this subsection.

Theorem 3 (Contractivity of Algorithm 2)**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g : \mathbb{R}^m \to \mathbb{R}$ be proper, convex, closed functions, and let g be L_g-smooth. Let $\phi_{\tau,\sigma}$ be as in* (14)*, and let $\epsilon > 0$ be arbitrary. Then, the fixed points of Algorithm 2 coincide with the solutions of* (1)*. Furthermore:*

(i) If $\tau\sigma\|A\|^2 + \frac{L_g}{2}\sigma < 1$ and either C1 or C2 hold, then Algorithm 2 is contractive in $\|\cdot\|_{\Psi_{\tau,\sigma}}$ with rate ρ, where $\Psi_{\tau,\sigma} = \Phi_{\tau,\sigma} + \mathrm{diag}(\gamma_x I_n, \mathbf{0}) \succ 0$,

$$\rho = \sqrt{1 - \frac{\min\{\gamma_x^2, \gamma_y^2\}}{\zeta_{\tau,\sigma} + \gamma_x}} < 1,$$

and γ_x and γ_y are as in Lemma 3 and Lemma 4, respectively; and

(ii) If $\tau\sigma\|A\|^2 + \frac{L_g}{2}\sigma \leq 1$, $\tau\sigma\|A\|^2(1+\epsilon)^2 < 1$ and C3 holds, then Algorithm 2 is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with rate

$$\rho = \frac{R_3'\zeta_{\tau,\sigma}}{\sqrt{{R_3'}^2\zeta_{\tau,\sigma}^2 + 1}} < 1,$$

where $R_3' > 0$ is independent of σ and τ. The best rate is obtained with $\tau = \sigma = \min\left\{\frac{\sqrt{L_g^2 + 16\|A\|^2} - L_g}{4\|A\|^2}, \frac{1}{(1+\epsilon)\|A\|}\right\}$.

Note that ϵ in Theorem 3(ii) is only used to account for the case $L_g = 0$ (in which instance, analogous considerations to Remark 1 hold), but $\epsilon = 0$ can otherwise be chosen. Deriving an expression for the optimal step sizes in Theorem 3(i) is cumbersome and not very insightful. Instead, let us focus on the novelty of Theorem 3. The step sizes bound $\tau\sigma\|A\|^2 + \frac{L_g}{2}\sigma < 1$ coincides with that of the Condat-Vũ method (Condat, 2013, Th. 3.1), (Lorenz and Pock, 2015, Th. 5). However, the linear rate of this algorithm has not been shown under C2 nor C3. Under C3, linear convergence was only studied for fully explicit methods (that use ∇f and ∇g) (Kovalev et al., 2022), while it might be convenient to employ the proximal operator of g, when available in closed form. On the other hand, known methods to achieve linear convergence under C2 (without further conditions) require both f and g to be prox-friendly, i.e., their prox operator can be efficiently computed (Wang and Xiao, 2017; Sadiev et al., 2022), in contrast to Algorithm 2. For instance, let us consider again the optimization problem in (4). Its Lagrangian is $\mathcal{L}(x,y) = -x^\top b + \iota_{\mathbb{R}^n_{\geq 0}}(x) + y^\top Ax - g(y)$; its saddle-point operator is $F(x,y) = (\mathrm{N}_{\mathbb{R}^n_{\geq 0}}(x) + A^\top y - b, \nabla g(y) - Ax)$, with $\mathrm{N}_{\mathbb{R}^m_{\geq 0}}$ the normal cone of the positive orthant, and Algorithm 2 reduces to

$$x^{k+1} = \mathrm{proj}_{\mathbb{R}^n_{\geq 0}}\left(x^k - \tau(A^\top y^k - b)\right) \tag{20}$$

$$y^{k+1} = y^k - \sigma\left(\nabla g(y^k) - A(2x^{k+1} - x^k)\right). \tag{21}$$

Theorem 3 ensures the contractivity (hence, Q-linear convergence) of this iteration under C2, i.e., if g is strongly convex and $A^\top$ is full-row rank. Linear convergence in this setup was only studied for equality constrained problems (Cisneros-Velarde et al., 2022), or by resorting to augmented Lagrangian formulations (Su et al., 2019), due to the difficulty of dealing with nonsmoothness, and of reconciling properties in weighted spaces with the unweighted projection $\mathrm{proj}_{\mathbb{R}^n_{\geq 0}}$ (Qu and Li, 2019). Theorem 3 circumvents these difficulties.

3.3 Preconditioned PDG method

Finally, we consider the method in (17), recalled in Algorithm 3. The iteration requires f and g to be smooth, but no proximal operator evaluations. This is often desirable to enable parallel and scalable computation, and to facilitate the adaptation to stochastic, mini-batch, and zeroth-order scenarios, for instance in empirical risk minimization with linear predictors (Xiao et al., 2019). If $\theta = -1$, Algorithm 3 retrieves the PDG method in (13); if $\theta = 0$, it retrieves the incremental PDG method. In both cases, convergence is not guaranteed without additional assumptions: an example where both methods fail is $f = g = 0$ and $A = 1$, which satisfies C3. Here, we again focus only on the case $\theta = 1$. This choice has the advantage of ensuring convergence for sufficiently small step sizes τ, σ, without strong convexity conditions; and it is the only value of θ that casts Algorithm 3 as a forward-backward method. In the following, we show that the choice $\theta = 1$ also suffices to ensure contractivity, even under C3.

Algorithm 3 Preconditioned PDG method

Require: $x^0 \in \mathbb{R}^n, y^0 \in \mathbb{R}^m$, step sizes $\tau, \sigma > 0$, $\theta = 1$
1: **for** $k = 0, 1, \cdots$ **do**
2: $\quad x^{k+1} = x^k - \tau \left(\nabla f(x^k) + A^\top y^k\right)$
3: $\quad y^{k+1} = y^k - \sigma \left(\nabla g(y^k) - A\left((1+\theta)x^{k+1} - \theta x^k\right)\right)$
4: **end for**

In particular, Algorithm 3 is derived as in (11) with $F_{\mathrm{b}}(\omega) = (A^\top y, -Ax)$, $F_{\mathrm{f}}(\omega) = (\nabla f(x), \nabla g(y))$ and $\Phi_{\tau,\sigma}$ as in (14), see Appendix B.3. By refining Lemma 3 and Lemma 1 to the case of Algorithm 3, it is easy to show that, under either C1, C2, or C3, condition A1 in Proposition 2 holds with $\Psi_{\mathrm{b}} = \mathbf{0}$, $\Psi_{\mathrm{b}} \not\succcurlyeq 0$, $\Psi_{\mathrm{b}} \succ 0$ respectively (we say that $\mathcal{B}$ is nonexpansive, partially contractive, contractive, respectively). The following lemma provides instead the contractivity properties for the forward step $\mathcal{F}$, necessary for our main result.

Lemma 5. *Let $F_{\mathrm{f}}(\omega) = (\nabla f(x), \nabla g(x))$. Select $\nu > 0$ arbitrary, $0 < \tau \leq \frac{2}{L_f + 2\|A\|\nu}$, and $0 < \sigma \leq \frac{2}{L_g + 2\|A\|\frac{1}{\nu}}$. Let*

$$\beta_x = \begin{cases} 2\mu_f - \frac{\tau}{1-\tau\|A\|\nu}\mu_f^2 & \text{if } \tau \leq \frac{2}{L_f+\mu_f+2\|A\|\nu} \\ 2L_f - \frac{\tau}{1-\tau\|A\|\nu}L_f^2 & \text{if } \tau > \frac{2}{L_f+\mu_f+2\|A\|\nu} \end{cases}$$

$$\beta_y = \begin{cases} 2\mu_g - \frac{\sigma}{1-\sigma\|A\|\frac{1}{\nu}}\mu_g^2 & \text{if } \sigma \leq \frac{2}{L_g+\mu_g+2\|A\|\frac{1}{\nu}} \\ 2L_g - \frac{\sigma}{1-\sigma\|A\|\frac{1}{\nu}}L_g^2 & \text{if } \sigma > \frac{2}{L_g+\mu_g+2\|A\|\frac{1}{\nu}} \end{cases}$$

Then $\beta_x > 0$ whenever $\mu_f > 0$ and $\tau < \frac{2}{L_f+2\|A\|\nu}$, and $\beta_f = 0$ otherwise. Likewise, $\beta_y > 0$ whenever $\mu_g > 0$ and $\sigma < \frac{2}{L_g+2\|A\|\frac{1}{\nu}}$, and $\beta_y = 0$ otherwise. Furthermore, A2 *in Proposition 2 holds with $\Psi_{\mathrm{f}} := \mathrm{diag}(\beta_x I_n, \beta_y I_m)$.*

Theorem 4 (Contractivity of Algorithm 3)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ be proper, convex, closed and smooth functions. Let $\Phi_{\tau,\sigma}$ be as in* (14)*, and let $\epsilon, \nu > 0$ be arbitrary. Then, the fixed points of Algorithm 3 coincide with the solutions of* (1)*. Furthermore:*

(i) If C1 holds and $\tau < \frac{2}{L_f+2\|A\|\nu}$, $\sigma < \frac{2}{L_g+2\|A\|\frac{1}{\nu}}$, then Algorithm 3 is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with rate

$$\rho = 1 - \min\left\{\frac{\beta_x \tau}{1+\tau\|A\|\nu}, \frac{\beta_y \sigma}{1+\sigma\|A\|\frac{1}{\nu}}\right\} < 1,$$

with β_x, β_y as in Lemma 5;

(ii) If C2 holds and $\tau \leq \frac{2}{L_f+2\|A\|\nu}$, $\sigma < \frac{2}{L_g+2\|A\|\frac{1}{\nu}}$, then Algorithm 3 is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}+\Psi_{\mathrm{b}}}$, $\Psi_{\mathrm{b}} = \mathrm{diag}\left(\frac{\mu_A}{\zeta_{\tau,\sigma}} I_n, \mathbf{0}\right)$, with rate

$$\rho = \sqrt{1 - \frac{\min\{\mu_A, \zeta_{\tau,\sigma}\beta_y\}}{\zeta_{\tau,\sigma}^2 + \mu_A\zeta_{\tau,\sigma}}} < 1,$$

with β_y as in Lemma 5; and

(iii) If C3 holds and $\tau \leq \frac{2}{L_f+2\|A\|\nu}$, $\sigma \leq \frac{2}{L_g+2\|A\|\frac{1}{\nu}}$, $\tau\sigma\|A\|^2(1+\epsilon)^2 \leq 1$, then Algorithm 3 is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with rate

$$\rho = \frac{\zeta_{\tau,\sigma}}{\sqrt{\mu_A + \zeta_{\tau,\sigma}^2}} < 1.$$

The best rate is obtained with $\tau = \sigma = \min\left\{\frac{1}{(1+\epsilon)\|A\|}, \frac{2}{L_f+2\|A\|\bar{\nu}}\right\}$, where $\bar{\nu} = \frac{L_g - L_f + \sqrt{(L_f-L_g)^2+16\|A\|^2}}{4\|A\|}$.

For $A = \mathbf{0}$, Theorem 4 retrieves the usual step sizes bounds for the gradient method and, under C1, the usual contractivity rate $\rho = \max\{|1-\tau\mu_f|, |1-\tau L_f|, |1-\sigma\mu_g|, |1-\sigma L_g|\}$ (see Proposition 5 in Appendix C.1). To the best of our knowledge, the work by Kovalev et al. (2022) is the only one to address linear convergence to solutions of (1) under C3 (and no further assumptions). The accelerated gradient method in (Kovalev et al., 2022) is optimal (i.e., it achieves the lower bound complexity) for the case that f and g are strongly convex, but not under C2 or C3. For example, when f and g are affine and A is invertible, both our Theorem 4 and Theorem 1 in Kovalev et al. (2022) guarantee the suboptimal rate $\rho = 1 - \Theta\left(\frac{\mu_A}{\|A\|^2}\right)$, where $\Theta(\cdot)$ denotes an asymptotic order, (for the optimal step sizes, and $\epsilon = 1$; as before, $\epsilon = 0$ can be chosen if instead either $L_g \neq 0$ or $L_f \neq 0$).

4 EXTENSIONS

Our contractivity certificates pave the way for several extensions of our technical results. We conclude this paper by discussing some of these extensions.

To start, acceleration schemes (e.g., overrelaxation and momentum (Iutzeler and Hendrickx, 2019), Anderson acceleration (Evans et al., 2020)) can directly be applied to Algorithms 1 to 3, given their interpretation as forward-backward (or proximal-point) methods and their contractivity. In some cases, improved theoretical rates can also be guaranteed (Evans et al., 2020).

In addition, contractivity implies important robustness properties, in case of inexact updates, delays, or problems with streaming data (see Appendix A.2). The operator-theoretic derivation also enables exploiting the extensive theory developed for nonexpansive and averaged operators (Bauschke and Combettes, 2017), in particular in case of inexact relaxations (Combettes, 2001). These observations should facilitate the convergence analysis for noisy and stochastic versions of our algorithms. We leave these generalizations for future work.

Our approach is not limited to the three algorithms studied nor to the forward-backward scheme in (11), but can be tailored to several problem formulations and serve as the basis for the development of fixed-point methods for many machine learning tasks. For example, we can show that the standard PDG method in (13) is contractive under C2; due to space limitations, we defer this result in Appendix D. The analysis is based on a weighted monotonicity argument, and it simplifies and extends the well-known work by Du and Hu (2019). In Appendix D we also design a novel algorithm to solve (1), that is contractive in the unweighted Euclidean norm.

Finally, let us note that the novel inverse-Lipschitz properties proven in Lemma 1 and Lemma 2 are algorithm independent, and of general interest. We believe that the inverse-Lipschitz condition has not received sufficient attention in the saddle-point literature; it would be interesting to study whether our techniques related to this property extend to general (convex-concave) saddle-point problems with nonlinear coupling.

References

Alghunaim, S. A. and Sayed, A. H. (2020). Linear convergence of primal–dual gradient methods and their performance in distributed optimization. *Automatica*, 117:109003.

Arrow, K. J., Hurwicz, L., and Uzawa, H. (1958). *Studies in Linear and Nonlinear Programming*, volume II. Stanford University Press, Stanford.

Bach, F., Mairal, J., and Ponce, J. (2008). Convex sparse matrix factorizations. arXiv:0812.1869.

Balamurugan, P. and Bach, F. (2016). Stochastic variance reduction methods for saddle-point problems. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 1416–1424.

Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3.

Banert, S., Upadhyaya, M., and Giselsson, P. (2023). The Chambolle–Pock method converges weakly with $\theta > 1/2$ and $\tau\sigma\|l\|^2 < 4/(1+2\theta)$. arXiv:2309.03998.

Bauschke, H. H. and Combettes, P. L. (2017). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer International Publishing, Cham.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press.

Bianchi, M., Belgioioso, G., and Grammatico, S. (2022). Fast generalized Nash equilibrium seeking under partial-decision information. *Automatica*, 136:110080.

Bredies, K., Chenchene, E., Lorenz, D. A., and Naldi, E. (2022). Degenerate preconditioned proximal point algorithms. *SIAM Journal on Optimization*, 32(3):2376–2401.

Bullo, F. (2024). *Contraction Theory for Dynamical Systems*. Kindle Direct Publishing, 1.2 edition.

Chambolle, A. and Pock, T. (2011). A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145.

Chambolle, A. and Pock, T. (2016a). An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319.

Chambolle, A. and Pock, T. (2016b). On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287.

Chen, G. H.-G. and Rockafellar, R. T. (1997). Convergence Rates in Forward–Backward Splitting. *SIAM Journal on Optimization*, 7(2):421–444.

Cisneros-Velarde, P., Jafarpour, S., and Bullo, F. (2022). A contraction analysis of primal-dual dynamics in distributed and time-varying implementations. *IEEE Transactions on Automatic Control*, 67(7):3560–3566.

Combettes, P. L. (2001). Quasi-Fejérian analysis of some optimization algorithms. In *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, volume 8 of *Studies in Computational Mathematics*, pages 115–152. Elsevier.

Combettes, P. L., Condat, L., Pesquet, J.-C., and Vũ, B. C. (2014). A forward-backward view of some primal-dual optimization methods in image recovery. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4141–4145.

Condat, L. (2013). A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479.

Condat, L., Kitahara, D., Contreras, A., and Hirabayashi, A. (2023). Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists. *SIAM Review*, 65(2):375–435.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2018). Training GANs with optimism. In *International Conference on Learning Representations*.

Davydov, A., Centorrino, V., Gokhale, A., Russo, G., and Bullo, F. (2025). Time-varying convex optimization: A contraction and equilibrium tracking approach. arXiv:2305.15595.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. (2017). Stochastic Variance Reduction Methods for Policy Evaluation. arXiv:1702.07944.

Du, S. S. and Hu, W. (2019). Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, pages 196–205.

Evans, C., Pollock, S., Rebholz, L. G., and Xiao, M. (2020). A proof that Anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically). *SIAM Journal on Numerical Analysis*, 58(1):788–810.

Fan, Q., Jiao, Y., and Lu, X. (2014). A Primal Dual Active Set Algorithm With Continuation for Compressed Sensing. *IEEE Transactions on Signal Processing*, 62(23):6276–6285.

Farina, G., Ling, C. K., Fang, F., and Sandholm, T. (2019). Correlation in Extensive-Form Games: Saddle-Point Formulation and Benchmarks. arXiv:1905.12564.

Gadjov, D. and Pavel, L. (2023). On the exact convergence to Nash equilibrium in hypomonotone regimes under full and partial-decision information. *IEEE Transactions on Automatic Control*, 68(8):4539–4553.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2019). A variational inequality perspective on generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019*.

Giselsson, P. and Moursi, W. M. (2021). On compositions of special cases of Lipschitz continuous operators. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):25.

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

Heise, P., Klose, S., Jensen, B., and Knoll, A. (2013). PM-Huber: Patchmatch with Huber regularization for stereo matching. In *2013 IEEE International Conference on Computer Vision*, pages 2360–2367.

Iutzeler, F. and Hendrickx, J. M. (2019). A generic online acceleration scheme for optimization algorithms via relaxation and inertia. *Optimization Methods and Software*, 34(2):383–405.

Jiang, B. and Schizas, I. D. (2021). Accelerated dual descent for network flow optimization. *IEEE Transactions on Signal Processing*, 69:1583–1597.

Jiang, F., Cai, X., and Han, D. (2023a). Inexact asymmetric forward-backward-adjoint splitting algorithms for saddle point problems. *Numerical Algorithms*, 94(1):479–509.

Jiang, F., Cai, X., and Han, D. (2023b). Inexact asymmetric forward-backward-adjoint splitting algorithms for saddle point problems. *Numerical Algorithms*, 94(1):479–509.

Jiang, F., Wu, Z., Cai, X., and Zhang, H. (2022). Unified linear convergence of first-order primal-dual algorithms for saddle point problems. *Optimization Letters*, 16(6):1675–1700.

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

Kovalev, D., Gasnikov, A., and Richtarik, P. (2022). Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. In *Advances in Neural Information Processing Systems*, volume 35, pages 21725–21737.

Kovalev, D., Salim, A., and Richtarik, P. (2020). Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 18342–18352.

Lei, Q., Yen, I. E.-H., Wu, C.-y., Dhillon, I. S., and Ravikumar, P. (2017). Doubly greedy primal-dual coordinate descent for sparse empirical risk minimization. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2034–2042.

Liang, T. and Stokes, J. (2019). Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, volume 89, pages 907–915.

Lorenz, D. A. and Pock, T. (2015). An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325.

Lorenz, D. A. and Schneppe, F. (2023). Chambolle–Pock's primal-dual method with mismatched adjoint. *Applied Mathematics & Optimization*, 87(2):22.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020). A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108, pages 1497–1507.

Monfort, A., Holly, A., and Gourieroux, C. (1982). Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, 50(1):63–80.

Moslem Zamani, H. A. and de Klerk, E. (2024). Convergence rate analysis of the gradient descent–ascent method for convex–concave saddle-point problems. *Optimization Methods and Software*, 0(0):1–23.

Nikolova, M. and Ng, M. K. (2005). Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific Computing*, 27(3):937–966.

O'Connor, D. and Vandenberghe, L. (2020). On the equivalence of the primal-dual hybrid gradient method and Douglas–Rachford splitting. *Mathematical Programming*, 179(1):85–108.

Qu, G. and Li, N. (2019). On the exponential stability of primal-dual gradient dynamics. *IEEE Control Systems Letters*, 3(1):43–48.

Qureshi, M. I. and Khan, U. A. (2023). Distributed saddle point problems for strongly concave-convex functions. *IEEE Transactions on Signal and Information Processing over Networks*, 9:679–690.

Rawlings, J. B., Mayne, D. Q., Diehl, M., et al. (2017). *Model predictive control: theory, computation, and design*, volume 2. Nob Hill Publishing Madison, WI.

Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898.

Ryu, E. K., Hannah, R., and Yin, W. (2022). Scaled relative graphs: nonexpansive operators via 2D Euclidean geometry. *Mathematical Programming*, 194(1):569–619.

Sadiev, A., Kovalev, D., and Richtarik, P. (2022). Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox. In *Advances in Neural Information Processing Systems*, volume 35, pages 21777–21791.

Salim, A., Condat, L., Kovalev, D., and Richtarik, P. (2022a). An optimal algorithm for strongly convex minimization under affine constraints. In *25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022)*, pages 4482–4498.

Salim, A., Condat, L., Mishchenko, K., and Richtárik, P. (2022b). Dualize, split, randomize: Toward fast nonsmooth optimization algorithms. *Journal of Optimization Theory and Applications*, 195(1):102–130.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shalev-Shwartz, S. and Zhang, T. (2014). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 64–72.

Su, Y., Shi, Y., and Sun, C. (2019). Contraction analysis on primal-dual gradient optimization. arXiv:1907.10171.

Tatarenko, T., Shi, W., and Nedić, A. (2021). Geometric convergence of gradient play algorithms for distributed Nash equilibrium seeking. *IEEE Transactions on Automatic Control*, 66(11):5342–5353.

Thekumparampil, K. K., He, N., and Oh, S. (2022). Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022)*, volume 151, pages 4281–4308.

Vũ, B. C. (2013). A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38.

Wang, J. and Xiao, L. (2017). Exploiting strong convexity from data with primal-dual first-order algorithms. In *34th International Conference on Machine Learning, ICML 2017*, volume 8, pages 3694–3702.

Wang, Y. and Li, J. (2020). Improved algorithms for convex-concave minimax optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 4800–4810.

Xiao, L., Yu, A. W., Lin, Q., and Chen, W. (2019). DSCOVR: randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *Journal of Machine Learning Research*, 20(1):1634–1691.

Xu, J., Tian, Y., Sun, Y., and Scutari, G. (2021). Distributed algorithms for composite optimization: Unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69:3555–3570.

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. (2004). Maximum margin clustering. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Zhang, G., Wang, Y., Lessard, L., and Grosse, R. B. (2022). Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 7659–7679.

Zhang, Y. and Lin, X. (2015). Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 353–361.

Zhu, Z., Chen, F., Zhang, J., and Wen, Z. (2023). A unified primal-dual algorithm framework for inequality constrained problems. *Journal of Scientific Computing*, 97(2):39.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] Conditions C1 to C3 and Theorems 2 to 4.
   (b) Complete proofs of all theoretical results. [Yes] See Appendix.
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Background Material

A.1 Uniqueness of primal-dual solutions

Lemma 6. *Let f and g be proper, convex, and closed functions. If any of C1, C2, or C3 hold, then the bilinear saddle point problem* (1) *has a unique solution* $(x^\star, y^\star)$.

Proof. We recall some properties of the Fenchel conjugate $g^*(v) = \sup\{\langle v, y\rangle - g(y) : y \in \mathbb{R}^m\}$ of a proper convex closed function g (Bauschke and Combettes, 2017, Ch.13, Th. 18.15): g^* is proper convex closed; if g is strongly convex, g^* is smooth (hence, $\mathrm{dom}(g) = \mathbb{R}^m$); if g is smooth, g^* is strongly convex; $\partial g^* = (\partial g)^{-1}$. Then, under any of the conditions C1, C2, or C3, the primal problem (2) has nonempty domain and is strongly convex (under C1, f is strongly convex; under C2 and C3, $g^* \circ A$ is), hence it admits a unique solution $x^\star \in \mathbb{R}^n$. Furthermore, under either condition C1, C2, or C3, we have that $\mathrm{dom}(g^*) \cap A(\mathrm{dom}(f)) \neq \varnothing$ (under C1 or C2, $\mathrm{dom}(g^*) = \mathbb{R}^m$ and f is proper; under C3, $A(\mathrm{dom}\, f) = \mathbb{R}^n$ and g^* is proper). Therefore, we can apply Theorem 19.1 in Bauschke and Combettes (2017); in particular, strong duality holds, and the solution to (2) is the unique vector $x^\star$ for which there exists $y^\star \in \mathbb{R}^m$ such that (Bauschke and Combettes, 2017, Theorem 19.1(ii))

$$\mathbf{0}_{n+m} \in F(x^\star, y^\star) := \begin{bmatrix} \partial f(x^\star) + A^\top y^\star \\ \partial g(y^\star) - Ax^\star \end{bmatrix}, \tag{22}$$

namely, such that $(x^\star, y^\star)$ solves (1). Hence, a solution $(x^\star, y^\star)$ to (1) must exist. Furthermore, the dual solution $y^\star$ is also unique, as, given the solution $x^\star$ of (2): if C1 or C2 hold, then ∂g is strongly monotone, hence the second inclusion in (22), $\mathbf{0} \in \partial g(y^\star) - Ax^\star$, has a unique solution $y^\star$ (Bauschke and Combettes, 2017, Ex. 22.12); if C3 holds, then the first inclusion in (22), $\mathbf{0} \in \nabla f(x^\star) + A^\top y^\star$, has a unique solution $y^\star$, since A is invertible. $\square$

That the conditions C1, C2, C3 are tight (i.e., that uniqueness of primal-dual solutions cannot be ensured if any of the subconditions is relaxed) can be verified via simple counterexamples. Consider the following cases:

(I) $n = 1$, $m = 2$, $f(x) = 0$, $g(y) = 0$, $A = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$;

(II) $n = m = 1$, $f(x) = 0$, $g = \iota_{[0,1]}(y) + y^2$, $A = 1$ (where $\iota_{[0,1]}$ is the indicator function of the interval $[0, 1]$);

(III) $n = m = 1$, $f(x) = 0$, $g(y) = y^2$, $A = 0$.

Recall the subconditions: C1(a): f is strongly convex (symmetrically, C1(b): g is strongly convex); C2(a) g is strongly convex; C2(b): g is smooth; C2(c) A is full column rank; C3(a): g is smooth (symmetrically, C3(b) f is smooth); C3(c): A is full row rank (symmetrically, C3(d): A is full column rank).

Case (III) violates C1(a) (but satisfies the other subconditions in C1). Case (I) violates C2(a). Case (II) violates C2(b). Case (III) violates C2(c). Case (II) violates C3(a). Case (I) violates C3(c).

It is easy to see that in all cases (I), (II), and (III), multiple primal-dual solutions exist for the inclusion in (22), hence for the saddle point problem in (1) (in which case, contractivity for *any* algorithm is excluded). Therefore, conditions C1, C2, C3 are tight.

A.2 Contractivity versus Q-linear convergence

We illustrate the difference between contractivity and Q-linear convergence. Let us consider an iteration $\omega^{k+1} = \mathcal{A}(\omega^k)$, where $\mathcal{A} : \mathbb{R}^q \to \mathbb{R}^q$ has a unique fixed point, i.e., there is a unique $\omega^\star \in \mathbb{R}^q$ such that $\mathcal{A}(\omega^\star) = \omega^\star$. The iteration is contractive (in $\|\cdot\|$) with rate $0 < \rho < 1$ if, for all $\omega, \omega' \in \mathbb{R}^q$, it holds that $\|\mathcal{A}(\omega) - \mathcal{A}(w')\| \leq \rho\|\omega - \omega'\|$, namely any two trajectories of the algorithm converge linearly to each other. The iteration is globally Q-linearly convergent with rate $0 < \rho < 1$ if, for all $\omega \in \mathbb{R}^q$, it holds that $\|\mathcal{A}(\omega) - \mathcal{A}(w^\star)\| \leq \rho\|\omega - \omega^\star\|$, therefore the iterates converge geometrically to $\omega^\star$.

Clearly, contractivity implies Q-linear convergence, by taking $\omega' = \omega^\star$. Conversely, there are many iterations that are Q-linearly convergent, but not contractive. For instance, the forward method for restricted strongly monotone operators (Tatarenko et al., 2021, Theorem 7), or the scalar iteration

$$\omega^{k+1} = \mathcal{A}(\omega^k) = \rho \sin(\omega)\omega, \tag{23}$$

with $0 < \rho < 1$,which converges Q-linearly to $\omega^\star = 0$ with rate ρ, but is not contractive. The gap between contractivity and Q-linear convergence is significant in terms of properties, proof strategy, and practical relevance.

Properties: Contractive algorithms enjoy superior robustness properties compared to Q-linear convergent iterations. For instance, contractivity ensures algorithm stability, in the sense that small changes in the initialization result in small changes in the execution path. This is not the case for Q-linearly convergent algorithms, as it can be checked on (23). As another example, consider the perturbed iteration $\omega^{k+1} = \mathcal{A}(\omega^k) + d(\omega^k)$. If $\mathcal{A}$ is ρ contractive, the error $d(\omega)$ is L_d-Lipschitz, and $\rho + L_d < 1$, then the perturbed iteration is still contractive, so it converges to a (perturbed) fixed point. This robustness cannot be guaranteed for Q-linearly convergent iterations. For instance, consider the perturbed version of (23), $\omega^{k+1} = \mathcal{A}(\omega^k) + \bar{d} = \rho \sin(\omega^k)\omega^k + \bar{d}$, where $\bar{d}$ is a constant independent of ω^k (so $L_d = 0$). For large enough $\bar{d} > 0$, the iteration has multiple fixed-points and it fails to converge, oscillating indefinitely. This would not happen for a contractive iteration. Finally, an important property of contractive iterations is modularity: the composition of contractive operators is contractive. At the contrary, if the iterations $\omega^{k+1} = \mathcal{A}(\omega^k)$ and $\omega^{k+1} = \mathcal{B}(\omega^k)$ are only Q-linearly convergent, the iteration $\omega^{k+1} = \mathcal{A} \circ \mathcal{B}(\omega^k)$ needs not converge.

Proof strategy: Proving contractivity is generally more involved than proving Q-linear convergence. One reason is that to show $||\mathcal{A}(w) - \mathcal{A}(w^\star)|| \leq \rho||\omega - \omega^\star||$, one can exploit the fact that $\omega^\star$ is a fixed point of $\mathcal{A}$. Of course, it is not always possible to extend Q-linear convergence arguments to contractivity, as there are algorithms that are Q-linearly convergent but not contractive, such as (23). More importantly, even if an iteration is contractive, a Lyapunov function that proves Q-linear convergence for an algorithm may not be suitable for demonstrating its contractivity, potentially requiring a different Lyapunov function.

Practical relevance: Because of its superior properties, contractivity plays an important role when an algorithm is used as a basis for the development of more complex methods. The robustness properties of contractive algorithms can be pivotal in studying perturbed iterations, or time-varying problems, such as those in learning with data streams; see the recent work by Davydov et al. (2025) for a continuous-time perspective. As another example, Xu et al. (2021) exploit the modularity of contractive iterations (specifically, of gradient descent) for the analysis of composite methods in distributed optimization (Xu et al., 2021, 7, Prop. 9); this derivation would not be possible based on Q-linear convergence guarantees alone. Finally, contractivity is required to guarantee convergence of general modification schemes, such as Anderson acceleration (Evans et al., 2020, Asm. 3.2).

A.3 Properties of $\Phi_{\tau,\sigma}$

Proposition 3. *Let* $\Phi_{\tau,\sigma} = \begin{bmatrix} \frac{1}{\tau} I_n & -A^\top \\ -A & \frac{1}{\sigma} I_m \end{bmatrix}$ *as in* (14). *Then:*

i) $\Phi_{\tau,\sigma}$ *is symmetric and positive definite if* $\tau\sigma\|A\|^2 < 1$.

ii) If $\tau\sigma\|A\|^2 < 1$, *the maximum eigenvalue of* $\Phi_{\tau,\sigma}$ *can then be upper bounded as follows*

$$\lambda_{\max}(\Phi_{\tau,\sigma}) = \|\Phi_{\tau,\sigma}\| \leq \max\left\{\frac{1}{\tau}, \frac{1}{\sigma}\right\} + \|A\| =: \zeta_{\tau,\sigma}.$$

Proof. Proof of item i). By Schur complement, $\Phi_{\tau,\sigma}$ is positive definite if and only if $\frac{1}{\tau} I_n - \sigma A^\top A \succ 0$, which is implied by $\tau\sigma\|A\|^2 < 1$.

Proof of item ii). Since $\Phi_{\tau,\sigma}$ is symmetric and positive semidefinite for $\tau\sigma\|A\|^2 \leq 1$, we have

$$\begin{aligned}
\lambda_{\max}(\Phi_{\tau,\sigma}) &= \|\Phi_{\tau,\sigma}\| \\
&= \| \begin{bmatrix} \frac{1}{\tau} I_n & 0 \\ 0 & \frac{1}{\sigma} I_m \end{bmatrix} + \begin{bmatrix} 0 & -A^\top \\ -A & 0 \end{bmatrix} \| \\
&\leq \max\left\{\frac{1}{\tau}, \frac{1}{\sigma}\right\} + \sqrt{\lambda_{\max}\left(\begin{bmatrix} A^\top A & 0 \\ 0 & AA^\top \end{bmatrix}\right)} \\
&= \max\left\{\frac{1}{\tau}, \frac{1}{\sigma}\right\} + \sqrt{\lambda_{\max}(AA^\top)} \\
&= \max\left\{\frac{1}{\tau}, \frac{1}{\sigma}\right\} + \|A\|.
\end{aligned}$$

□

A.4 Properties of inverse Lipschitz operators

Lemma 7. *Let $\mathcal{A} : \mathbb{R}^q \rightrightarrows \mathbb{R}^q$ be $\frac{1}{L}$-inverse Lipschitz, i.e., $\|u - u'\| \geq \frac{1}{L}\|\omega - \omega'\|$ for all $(\omega, u), (\omega', u') \in \operatorname{gra}(\mathcal{A})$. Then $\mathcal{A}$ has at most one zero, i.e., there exists at most one $\omega^\star \in \mathbb{R}^q$ such that $\mathbf{0} \in \mathcal{A}(\omega^\star)$.*

Proof. The definition of inverse Lipschitz implies that, if $(\omega, \mathbf{0}), (\omega', \mathbf{0}) \in \operatorname{gra}(\mathcal{A})$, then $\|\mathbf{0} - \mathbf{0}\| \geq \frac{1}{L}\|\omega - \omega'\|$, and hence $\omega = \omega'$. □

Lemma 8. *Let $\mathcal{A} : \mathbb{R}^q \rightrightarrows \mathbb{R}^q$ be μ-strongly monotone. Then, $\mathcal{A}$ is μ-inverse Lipschitz.*

Proof. By definition of strong monotonicity and the Cauchy–Schwartz inequality, for any $(\omega, u), (\omega', u') \in \operatorname{gra}(\mathcal{A})$, $\mu\|\omega - \omega'\|^2 \leq \langle u - u' \mid \omega - \omega' \rangle \leq \|u - u'\|\|\omega - \omega' \ \|$, and hence $\|u - u'\| \geq \mu\|\omega - \omega'\|$. □

B Algorithms derivation

The three algorithms (15), (16) and (17), are derived from forward-backward splitting in (11). The preconditioning matrix $\Phi_{\tau,\sigma}$ is defined in (14), and we assume throughout that $\tau\sigma\|A\|^2 < 1$, so that $\Phi_{\tau,\sigma} \succ 0$. The forward-backward iteration in (11) can thus be reformulated as follows

$$\begin{aligned} & \omega^{k+1} = (\operatorname{Id} + \Phi_{\tau,\sigma}^{-1} F_{\mathrm{b}})^{-1} \circ (\operatorname{Id} - \Phi_{\tau,\sigma}^{-1} F_{\mathrm{f}})(\omega^k) \\ \overset{(i)}{\Leftrightarrow} \ & \left(\operatorname{Id} + \Phi_{\tau,\sigma}^{-1} F_{\mathrm{b}}\right)(\omega^{k+1}) \ni \left(\operatorname{Id} - \Phi_{\tau,\sigma}^{-1} F_{\mathrm{f}}\right)(\omega^k) \\ \overset{(ii)}{\Leftrightarrow} \ & \left(\Phi_{\tau,\sigma} + F_{\mathrm{b}}\right)(\omega^{k+1}) \ni \left(\Phi_{\tau,\sigma} - F_{\mathrm{f}}\right)(\omega^k), \end{aligned} \tag{24}$$

where (i) follows from the definition of inverse operator, and (ii) holds since $\Phi_{\tau,\sigma}$ is invertible.

Remark 2. *For either of Algorithms 1 to 3, the operators F_{f} and F_{b} are maximally monotone (Bauschke and Combettes, 2017, Cor. 25.5), as they all can be written as the sum of the subdifferential of a convex function (which is maximally monotone) and a linear skew symmetric operator (which is maximally monotone and with full domain). In particular, this implies that the backward step $\mathcal{B} = (\operatorname{Id} + \Phi_{\sigma,\tau}^{-1} F_{\mathrm{b}})^{-1}$ is always single-valued and has full domain (Bauschke and Combettes, 2017, Prop. 20.14, Prop. 23.8), even though F_{b} might be set-valued (hence the "$\ni$" rather than "$=$" sign in the first equivalence above).*

B.1 Algorithm 1 (Equation (15))

We choose $F_{\mathrm{f}} = \mathbf{0}$, $F_{\mathrm{b}} = F$ as described in Section 3.1. Explicitly writing (24) yields the following:

$$\begin{aligned} & \left(\Phi_{\tau,\sigma} + F_{\mathrm{b}}\right)(\omega^{k+1}) \ni \left(\Phi_{\tau,\sigma} - F_{\mathrm{f}}\right)(\omega^k) \\ \Leftrightarrow & \begin{bmatrix} \frac{1}{\tau}x^{k+1} + \partial f(x^{k+1}) + \cancel{A^\top y^{k+1}} - \cancel{A^\top y^{k+1}} \\ \frac{1}{\sigma}y^{k+1} - Ax^{k+1} - Ax^{k+1} \end{bmatrix} \ni \begin{bmatrix} \frac{1}{\tau}x^k - A^\top y^k \\ \frac{1}{\sigma}y^k - Ax^k \end{bmatrix} \\ \Leftrightarrow & \begin{bmatrix} x^{k+1} + \tau\partial f(x^{k+1}) \\ y^{k+1} + \sigma\partial g(y^{k+1}) \end{bmatrix} \ni \begin{bmatrix} x^k - \tau A^\top y^k \\ y^k + \sigma A\left(2x^{k+1} - x^k\right) \end{bmatrix} \\ \Leftrightarrow & \begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} \operatorname{prox}_{\tau f}\left(x^k - \tau A^\top y^k\right) \\ \operatorname{prox}_{\sigma g}\left(y^k + \sigma A\left(2x^{k+1} - x^k\right)\right), \end{bmatrix} \end{aligned}$$

where the last equivalence follows by definition of prox operator (we recall that for a proper closed convex function ψ, $\operatorname{prox}_\phi = (\operatorname{Id} + \partial\psi)^{-1}$ is single valued, see (9), although $\partial\psi$ might not be; hence "$=$" rather than "$\in$" is used in the last line).

Remark 3 (Design rationale)**.** *Informally speaking, convergence of the forward-backward method* (11) *typically requires only monotonicity for the operator F_{b}, but a stronger "cocoercivity" condition for the operator F_{f} (Bauschke and Combettes, 2017, Th. 26.14) (cocoercivity is for instance satisfied by gradients of smooth functions). For this reason, to ensure convergence even without strong convexity assumptions, we always place the skew symmetric operator $(A^\top y, -Ax)$, which is not cocoercive, in the backward step. On the downside, this complicates the computation of $\mathcal{B}$. To remedy, the preconditioning matrix $\Phi_{\tau,\sigma}$ is designed to make the system of inclusions above* *block triangular, i.e., to remove the term $A^\top y^{k+1}$ in the first inclusion. In this way, x^{k+1} does not depend on y^{k+1}. This ensures that the resulting iteration can be computed efficiently (provided that the proximal operators for the functions f and g are available).*

B.2 Algorithm 2 (Equation (16))

We choose $F_\mathrm{f}(\omega) = (\mathbf{0}, \nabla g(y))$ and $F_\mathrm{b}(\omega) = (\partial f(x) + A^\top y, -Ax)$ as described in Section 3.2. Then, we have from (24):

$$
\begin{aligned}
& \left(\Phi_{\tau,\sigma} + F_\mathrm{b}\right)(\omega^{k+1}) \ni \left(\Phi_{\tau,\sigma} - F_\mathrm{f}\right)(\omega^k) \\
\Leftrightarrow & \begin{bmatrix} \frac{1}{\tau}x^{k+1} + \partial f(x^{k+1}) + \cancel{A^\top y^{k+1}} - \cancel{A^\top y^{k+1}} \\ \frac{1}{\sigma}y^{k+1} - Ax^{k+1} - Ax^{k+1} \end{bmatrix} \ni \begin{bmatrix} \frac{1}{\tau}x^k - A^\top y^k \\ \frac{1}{\sigma}y^k - Ax^k - \nabla g(y^k) \end{bmatrix} \\
\Leftrightarrow & \begin{bmatrix} x^{k+1} + \tau \partial f(x^{k+1}) \\ y^{k+1} \end{bmatrix} \ni \begin{bmatrix} x^k - \tau A^\top y^k \\ y^k - \sigma\left(\nabla g(y^k) - A\left(2x^{k+1} - x^k\right)\right) \end{bmatrix} \\
\Leftrightarrow & \begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} \mathrm{prox}_{\tau f}\left(x^k - \tau A^\top y^k\right) \\ y^k - \sigma\left(\nabla g(y^k) - A\left(2x^{k+1} - x^k\right)\right), \end{bmatrix}
\end{aligned}
$$

which concludes the derivation of Algorithm 2.

B.3 Algorithm 3 (Equation (17))

We choose $F_\mathrm{b}(\omega) = (A^\top y, -Ax)$ and $F_\mathrm{f}(\omega) = (\nabla f(x), \nabla g(y))$ as given in Section 3.3. From (24), we have

$$
\begin{aligned}
& \left(\Phi_{\tau,\sigma} + F_\mathrm{b}\right)(\omega^{k+1}) \ni \left(\Phi_{\tau,\sigma} - F_\mathrm{f}\right)(\omega^k) \\
\Leftrightarrow & \begin{bmatrix} \frac{1}{\tau}x^{k+1} + \cancel{A^\top y^{k+1}} - \cancel{A^\top y^{k+1}} \\ \frac{1}{\sigma}y^{k+1} - Ax^{k+1} - Ax^{k+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\tau}x^k - A^\top y^k - \nabla f(x^k) \\ \frac{1}{\sigma}y^k - Ax^k - \nabla g(y^k) \end{bmatrix} \\
\Leftrightarrow & \begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} x^k - \tau\left(\nabla f(x^k) + A^\top y^k\right) \\ y^k - \sigma\left(\nabla g(y^k) - A\left(2x^{k+1} - x^k\right)\right), \end{bmatrix}
\end{aligned}
$$

where in the first equivalence we have equality since all operators are single-valued.

C Omitted Proofs

C.1 Proof of Proposition 1

We will also recall the case of strongly monotone operators.

Proposition 4 (Contractivity of Resolvents)**.** *Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximally monotone operator. Then its resolvent $\mathrm{J}_F = (\mathrm{Id} + F)^{-1}$ is everywhere defined and single-valued. Moreover:*

(i) If F is μ-strongly monotone, then the resolvent J_F is $\frac{1}{1+\mu}$-Lipschitz.

(ii) If F is $\frac{1}{L}$ inverse Lipschitz, then the resolvent J_F is $\frac{L}{\sqrt{L^2+1}}$-Lipschitz.

Proof. That the resolvent of a maximally monotone operator is full-domain and single valued is the well-known Minty's Theorem (Bauschke and Combettes, 2017, Prop. 23.8).

Proof of item i). The operator $\mathrm{Id} + \mathrm{F}$ is $(1+\mu)$ -strongly monotone, so its inverse is $\frac{1}{1+\mu}$-Lipschitz.

Proof of item ii). For any $x, x' \in \mathbb{R}^n$, let $u = \mathrm{J}_F(x)$ and $u' = \mathrm{J}_F(x')$, which implies $u + F(u) \ni x'$ and $u' + F(u') \ni x'$, by definition of inverse operator. Hence, for some $v \in F(u)$ and $v' \in F(u')$, we have

$$
\begin{aligned}
\|x - x'\|^2 &= \|u + v - u' - v'\|^2 \\
&= \|u - u'\|^2 + \|v - v'\|^2 + 2\langle u - u', v - v' \rangle \\
&\overset{(i)}{\geq} \|u - u'\|^2 + \frac{1}{L^2}\|u - u'\|^2 \\
&= \left(1 + \frac{1}{L^2}\right)\|u - u'\|^2,
\end{aligned}
$$

where monotonicity and the inverse Lipschitz property of A were used in (i). Taking a square root on both sides proves that $(\mathrm{Id}+F)^{-1}$ is Lipschitz with constant $\frac{L}{\sqrt{L^2+1}} < 1$. ☐

For future reference, let us also recall the following, well-known, result (we will prove a more general result in Theorem 4, hence we do not include a proof here).

Proposition 5 (Contractivity of Gradient Step). *Let $f : \mathbb{R}^n \to R$ be μ-strongly convex and L-smooth. Then, for any $0 < \alpha < \frac{2}{L}$, the operator $(\mathrm{Id} - \alpha\nabla f)$ is ρ-Lipschitz, with $\rho = \max\{|1-\alpha L|, |1-\alpha\mu|\}$.*

C.2 Proof of Proposition 2

For simplicity of notation, we use a slightly refined (but equivalent) version of Proposition 2 in some of our proofs. Clearly, Proposition 2 is retrieved from Proposition 6 by choosing $\rho_{\mathrm{f}} = \rho_{\mathrm{b}} = 1$.

Proposition 6. *Let $\Phi_{\tau,\sigma} \succ 0$ as in (14). Assume that there exists scalars $0 \le \rho_{\mathrm{f}}, \rho_{\mathrm{b}}, \gamma \le 1$ and matrices $\Psi_{\mathrm{b}} \succcurlyeq 0$, $\Psi_{\mathrm{f}} \succcurlyeq 0$, such that, for all $\omega, \omega' \in \mathbb{R}^{n+m}$,*

A1. $\|\mathcal{B}(\omega) - \mathcal{B}(\omega')\|^2_{\Phi_{\tau,\sigma}+\Psi_{\mathrm{b}}} \le \rho_{\mathrm{b}}^2 \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}}$

A2. $\|\mathcal{F}(\omega) - \mathcal{F}(\omega')\|^2_{\Phi_{\tau,\sigma}} \le \rho_{\mathrm{f}}^2 \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}-\Psi_{\mathrm{f}}}$

and $\Psi_{\mathrm{b}} + \Psi_{\mathrm{f}} \succcurlyeq \gamma(\Phi_{\tau,\sigma} + \Psi_{\mathrm{b}})$. Then, the forward-backward iteration in (11) is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}+\Psi_{\mathrm{b}}}$, with rate $\rho = \rho_{\mathrm{b}}\rho_{\mathrm{f}}\sqrt{1-\gamma}$.

Proof. Let $\omega, \omega' \in \mathbb{R}^{n+m}$. We have

$$\begin{aligned}
&\|\mathcal{B}\circ\mathcal{F}(\omega) - \mathcal{B}\circ\mathcal{F}(\omega')\|^2_{\Phi_{\tau,\sigma}+\Psi_b} \\
&\overset{(i)}{\le} \rho_{\mathrm{b}}^2 \|\mathcal{F}(\omega) - \mathcal{F}(\omega')\|^2_{\Phi_{\tau,\sigma}} \\
&\overset{(ii)}{\le} \rho_{\mathrm{b}}^2\rho_{\mathrm{f}}^2 \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}-\Psi_f} \\
&= \rho_{\mathrm{b}}^2\rho_{\mathrm{f}}^2 \left(\omega - \omega'\right)^\top \left(\Phi_{\tau,\sigma} - \Psi_f\right)\left(\omega - \omega'\right) \\
&\overset{(iii)}{\le} \rho_{\mathrm{b}}^2\rho_{\mathrm{f}}^2 \left(\omega - \omega'\right)^\top \left(\Phi_{\tau,\sigma} + \Psi_b - \gamma(\Phi_{\tau,\sigma} + \Psi_b)\right)\left(\omega - \omega'\right) \\
&= \rho_{\mathrm{b}}^2\rho_{\mathrm{f}}^2 (1-\gamma)\left(\omega - \omega'\right)^\top \left(\Phi_{\tau,\sigma} + \Psi_b\right)\left(\omega - \omega'\right) \\
&= \rho_{\mathrm{b}}^2\rho_{\mathrm{f}}^2 (1-\gamma)\|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}+\Psi_b},
\end{aligned}$$

where (i) follows from A1, (ii) from A2 and (iii) from $\Psi_{\mathrm{b}} + \Psi_{\mathrm{f}} \succcurlyeq \gamma(\Phi_{\tau,\sigma} + \Psi_{\mathrm{b}})$. Taking the square root on both sides concludes the proof. ☐

C.3 Proof of Lemma 1

The lemma assumes that C2 holds, hence g is strongly convex and smooth and the matrix A has full column rank. Define the following matrix $\Psi_\alpha \in \mathbb{R}^{(n+m)\times(n+m)}$, for $\alpha > 0$ small enough to be chosen,

$$\Psi_\alpha = \begin{bmatrix} I_n & 0 \\ -\alpha A & I_m \end{bmatrix}.$$

For any $\omega = (x, y), \omega' = (x', y') \in \mathbb{R}^{n+m}$, let $v \in F(\omega)$, $v' \in F(\omega')$, so that, for some $u \in \partial f(x)$ and $u' \in \partial f(x')$

$$\begin{aligned}
&\|v - v'\|\|\Psi_\alpha\|\|\omega - \omega'\| \\
&= \left\| \begin{bmatrix} u + A^\top y - (u' + A^\top y') \\ \nabla g(y) - Ax - (\nabla g(y') - Ax') \end{bmatrix} \right\| \left\| \begin{bmatrix} I_n & 0 \\ -\alpha A & I_m \end{bmatrix} \right\| \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\| \\
&\ge \left\| \begin{bmatrix} u + A^\top y - (u' + A^\top y') \\ \nabla g(y) - Ax - (\nabla g(y') - Ax') \end{bmatrix} \right\| \left\| \begin{bmatrix} x - x' \\ -\alpha A(x - x') + y - y' \end{bmatrix} \right\| \\
&\overset{(i)}{\ge} \left\langle \begin{bmatrix} u + A^\top y - (u' + A^\top y') \\ \nabla g(y) - Ax - (\nabla g(y') - Ax') \end{bmatrix}, \begin{bmatrix} x - x' \\ -\alpha A(x - x') + y - y' \end{bmatrix} \right\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \langle u-u', x-x'\rangle + \cancel{\left\langle A^\top(y-y'), x-x'\right\rangle} + \langle \nabla g(y)-\nabla g(y'), y-y'\rangle \\
&\quad -\alpha\langle \nabla g(y)-\nabla g(y'), A(x-x')\rangle + \alpha\|A(x-x')\|^2 - \cancel{\langle A(x-x'), y-y\rangle} \\
&\overset{(ii)}{\geq} \langle \nabla g(y)-\nabla g(y'), y-y'\rangle - \alpha\langle \nabla g(y)-\nabla g(y'), A(x-x')\rangle \\
&\quad +\mu_f\|x-x'\|^2 + \alpha\|A(x-x')\|^2 \\
&\overset{(iii)}{\geq} \mu_g\|y-y'\|^2 - \alpha L_g\|A\|\|x-x'\|\|y-y'\| \\
&\quad \mu_f\|x-x'\|^2 + \alpha\lambda_{\min}\left(A^\top A\right)\|x-x'\| \\
&= \begin{bmatrix}\|x-x'\| & \|y-y'\|\end{bmatrix} \underbrace{\begin{bmatrix}\mu_f+\alpha\lambda_{\min}\left(A^\top A\right) & -\frac{\alpha L_g\|A\|}{2} \\ -\frac{\alpha L_g\|A\|}{2} & \mu_g\end{bmatrix}}_{:=M_\alpha} \begin{bmatrix}\|x-x'\| \\ \|y-y'\|\end{bmatrix} \\
&\geq \lambda_{\min}\left(M_\alpha\right)\|\omega-\omega'\|^2,
\end{aligned}
$$

where (i) follows from the Cauchy–Schwarz inequality, (ii) from (strong) convexity of f (with $\mu_f = 0$ possibly) and (iii) from strong convexity and smoothness of g. By Sylvester's criterion, the matrix M_α is positive definite (hence $\lambda_{\min}(M_\alpha) > 0$) for α small enough, in particular if we choose

$$
\alpha < \frac{2\lambda_{\min}\left(A^\top A\right)\mu_g + 2\sqrt{\lambda_{\min}^2(A^\top A)\mu_g^2 + L_g^2\|A\|^2\mu_g\mu_f}}{L_g^2\|A\|^2}.
$$

Using that $\lambda_{\max}(\Psi_\alpha) = \|\Psi_\alpha\| \leq 1+\alpha\|A\|$ yields $\|v-v'\| \geq \frac{\lambda_{\min}(M_\alpha)}{1+\alpha\|A\|}\|\omega-\omega'\|$, which concludes the proof with $R_2 = \frac{1+\alpha\|A\|}{\lambda_{\min}(M_\alpha)} > 0$. □

C.4 Proof of Lemma 2

The idea is very similar to the proof of Lemma 1. Under C3, both functions f and g are smooth and the matrix A is invertible. Define the following matrix,$\Psi_\varepsilon \in \mathbb{R}^{2n\times 2n}$, for $\varepsilon > 0$ small enough to be chosen,

$$
\Psi_\varepsilon = \begin{bmatrix} I_n & \varepsilon A^\top \\ -\varepsilon A & I_n \end{bmatrix}.
$$

Then we have, for any $\omega = (x,y), \omega' = (x',y') \in \mathbb{R}^{2n}$,

$$
\begin{aligned}
&\|F(\omega)-F(\omega')\|\|\Psi_\varepsilon\|\|\omega-\omega'\| \\
&\geq \left\| \begin{bmatrix}\nabla f(x)-\nabla f(x')+A^\top(y-y') \\ \nabla g(y)-\nabla g(y')-A(x-x')\end{bmatrix} \right\| \left\| \begin{bmatrix} x-x'+\varepsilon A^\top(y-y') \\ y-y'-\varepsilon A(x-x')\end{bmatrix}\right\| \\
&\overset{(i)}{\geq} \left\langle \begin{bmatrix}\nabla f(x)-\nabla f(x')+A^\top(y-y') \\ \nabla g(y)-\nabla g(y')-A(x-x')\end{bmatrix}, \begin{bmatrix} x-x'+\varepsilon A^\top(y-y') \\ y-y'-\varepsilon A(x-x')\end{bmatrix}\right\rangle \\
&= \langle \nabla f(x)-\nabla f(x'), x-x'\rangle + \varepsilon\left\langle \nabla f(x)-\nabla f(x'), A^\top(y-y')\right\rangle + \varepsilon\|A^\top(y-y')\|^2 \\
&\quad + \langle \nabla g(y)-\nabla g(y'), y-y'\rangle - \varepsilon\langle \nabla g(y)-\nabla g(y'), A(x-x')\rangle + \varepsilon\|A(x-x')\|^2 \\
&\overset{(ii)}{\geq} \frac{1}{L_f}\|\nabla f(x)-\nabla f(x')\|^2 + \varepsilon\left\langle \nabla f(x)-\nabla f(x'), A^\top(y-y')\right\rangle + \varepsilon\|A^\top(y-y')\|^2 \\
&\quad + \frac{1}{L_g}\|\nabla g(y)-\nabla g(y')\|^2 - \varepsilon\langle \nabla g(y)-\nabla g(y'), A(x-x')\rangle + \varepsilon\|A(x-x')\|^2 \\
&\overset{(iii)}{\geq} \left(\frac{1}{L_f} - \frac{\|A\|\varepsilon\delta}{2}\right)\|\nabla f(x)-\nabla f(x')\|^2 + \varepsilon\left(\lambda_{\min}\left(AA^\top\right) - \frac{\|A\|}{2\delta}\right)\|y-y'\|^2 \\
&\quad + \left(\frac{1}{L_g} - \frac{\|A\|\varepsilon\delta}{2}\right)\|\nabla g(y)-\nabla g(y')\|^2 + \varepsilon\left(\lambda_{\min}\left(A^\top A\right) - \frac{\|A\|}{2\delta}\right)\|x-x'\|^2
\end{aligned}
$$

where (i) follows from Cauchy-Schwarz, (ii) from cocoercivity of ∇f and ∇g, (iii) from Young's inequality, for any $\delta > 0$. In particular, if we choose $\delta > \frac{\|A\|}{2\mu_A}$, $\varepsilon < \frac{2}{L\|A\|\delta}$, $L = \max\{L_f, L_g\}$ we obtain $\|F(\omega)-F(\omega')\|\|\Psi_\varepsilon\|\|\omega-\omega'\| \geq C(\epsilon,\delta)\|\omega-\omega'\|^2$, where

$$
C(\varepsilon,\delta) = \varepsilon\left(\mu_A - \frac{\|A\|}{2\delta}\right) > 0.
$$

Using $\|\Psi_\varepsilon\| \leq 1 + \varepsilon\|A\|$ concludes the proof, with $R_3 = \frac{1+\varepsilon\|A\|}{C(\varepsilon,\delta)} > 0$. ☐

C.5 Proof of Theorem 2

We first recall an additional result.

Lemma 9. *Let $F : \mathbb{R}^{n+m} \rightrightarrows \mathbb{R}^{n+m}$ be μ-strongly monotone in $\langle \cdot, \cdot \rangle$, with $\mu \geq 0$, and $\Phi_{\tau,\sigma} \succ 0$. Then $\Phi_{\tau,\sigma}^{-1} F$ is strongly monotone in $\langle \cdot, \cdot \rangle_{\Phi_{\tau,\sigma}}$ with parameter $\frac{\mu}{\lambda_{\max}(\Phi_{\tau,\sigma})}$, i.e., for any $(\omega, u), (\omega', u') \in \operatorname{gra}(\Phi_{\tau,\sigma}^{-1} F)$, it holds that $\langle u - u', \omega - \omega' \rangle_{\Phi_{\tau,\sigma}} \geq \frac{\mu}{\lambda_{\max}(\Phi_{\tau,\sigma})} \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}}$.*

Proof. Note that $(\omega, u) \in \operatorname{gra}(F)$ if and only if $(\omega, \Phi_{\tau,\sigma}^{-1} u) \in \operatorname{gra}(\Phi_{\tau,\sigma}^{-1} F)$. For any $(\omega, u), (\omega', u') \in \operatorname{gra} F$, we have

$$\begin{aligned}\left\langle \Phi_{\tau,\sigma}^{-1} u - \Phi_{\tau,\sigma}^{-1} u', \omega - \omega' \right\rangle_{\Phi_{\tau,\sigma}} &= \langle u - u', \omega - \omega' \rangle \\ &\geq \mu \|\omega - \omega'\|^2 \\ &\geq \frac{\mu}{\lambda_{\max}(\Phi_{\tau,\sigma})} \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}},\end{aligned}$$

where the last inequality follows from equivalence of norms. ☐

For any $\tau > 0, \sigma > 0$, the fixed points of Algorithm 1 are the vectors $(x^\star, \omega^\star)$ that satisfy

$$\begin{aligned} x^\star &= \operatorname{prox}_{\tau f}\left(x^\star - \tau A^\top y^\star\right) \\ y^\star &= \operatorname{prox}_{\sigma g}\left(y^\star + \sigma A(2x^\star - x^\star)\right), \end{aligned}$$

or, equivalently, by definition of proximal operator,

$$\begin{aligned} x^\star + \tau \partial f(z^\star) &\ni x^\star - \tau A^\top y^\star \\ y^\star + \sigma \partial g &\ni y^\star + \sigma A x^\star, \end{aligned}$$

equivalently, $\mathbf{0} \in F(x^\star, y^\star)$, equivalently the solutions of (1).

Furthermore, whenever $\sigma\tau\|A\|^2 < 1$ (so that $\Phi_{\tau,\sigma} \succ 0$), Algorithm 1 can be recast as the forward-backward iteration in (11) (see Appendix B.1), with $F_{\mathrm{f}} = \mathbf{0}$, $F_{\mathrm{b}} = F$. Since $F_{\mathrm{f}} = \mathbf{0}$, then the forward-step $\mathcal{F} = \operatorname{Id}$ satisfies trivially A2 in Proposition 6 with $\Psi_{\mathrm{f}} = \mathbf{0}$, $\rho_{\mathrm{f}} = 1$ (namely, $\mathcal{F}$ is nonexpansive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$). It is left to show that $\mathcal{B}$ is contractive, which is done in the following, for each of the conditions C1, C2, C3.

C.5.1 Item (i): Algorithm 1 under C1

We start by an auxiliary result. Let $\tilde{\rho} := \min\{\mu_f \tau, \mu_g \sigma, \kappa\}$. We prove that

$$\begin{bmatrix} \mu_f I & 0 \\ 0 & \mu_g I \end{bmatrix} - \tilde{\rho}\Phi_{\tau,\sigma} = \begin{bmatrix} \left(\mu_f - \tilde{\rho}\frac{1}{\tau}\right) I & \tilde{\rho} A^\top \\ \tilde{\rho} A & \left(\mu_g - \tilde{\rho}\frac{1}{\sigma}\right) I \end{bmatrix} \succcurlyeq 0. \tag{25}$$

If $A = \mathbf{0}$, this is obvious. If $\|A\| \neq 0$, then either $\tilde{\rho} < \mu_f \tau$ or $\tilde{\rho} < \mu_g \sigma$, because $\kappa = \frac{\mu_f \tau}{1+\sqrt{\tau\sigma\|A\|^2}} < \mu_g \tau$ if $\mu_f \tau = \mu_g \sigma$. Let us assume without loss of generality that $\tilde{\rho} < \mu_f \tau$. Then, by Schur complement, (25) is equivalent to

$$\left(\mu_g - \tilde{\rho}\frac{1}{\sigma}\right) I - \tilde{\rho}^2 A \left[\left(\mu_f - \tilde{\rho}\frac{1}{\tau}\right) I\right]^{-1} A^\top \succcurlyeq 0,$$

which is implied by (since $AA^\top \succcurlyeq 0$)

$$\begin{aligned} &\left(\mu_f - \tilde{\rho}\frac{1}{\tau}\right)\left(\mu_g - \tilde{\rho}\frac{1}{\sigma}\right) - \tilde{\rho}^2\|A\|^2 \geq 0 \\ \iff \quad &\tilde{\rho}^2(1 - \tau\sigma\|A\|^2) - \rho(\mu_f \tau + \mu_g \sigma) + \mu_f \mu_g \tau\sigma \geq 0. \end{aligned}$$

In turn, by solving the second-order inequality, we obtain that (25) holds if $\tilde{\rho} \leq \kappa$ (note that $1 - \tau\sigma\|A\|^2 > 0$ and the determinant of the second order inequality is $(\mu_f \tau + \mu_g \sigma)^2 - 4(1 - \tau\sigma\|A\|^2)(\mu_f \mu_g \tau\sigma) \geq (\mu_f \tau - \mu_g \sigma)^2 \geq 0$), which proves (25). We further note that $\kappa > 0$, because $(\mu_f \tau - \mu_g \sigma)^2 + 4\|A\|^2 \mu_f \mu_g \tau^2 \sigma^2 < (\mu_f \tau + \mu_g \sigma)^2$ whenever $\tau\sigma\|A\|^2 < 1$.

Now, for $\omega, \omega' \in \mathbb{R}^{n+m}$, let $v = (I + \Phi_{\tau,\sigma}^{-1} F)^{-1}\omega$ and $v' = (I + \Phi_{\tau,\sigma}^{-1} F)^{-1}\omega'$, which implies $\omega = v + \Phi_{\tau,\sigma}^{-1} z$ and $\omega' = v' + \Phi_{\tau,\sigma}^{-1} z'$ for some $z \in Fv$ and $z' \in Fv'$, by definition of inverse operator. Therefore, by strong convexity, definition of F, and letting $v = (v_x, v_y)$, $v' = (v'_x, v'_y)$, we have

$$\begin{aligned}
&\|\omega - \omega'\|_{\Phi_{\tau,\sigma}} \|v - v'\|_{\Phi_{\tau,\sigma}} \\
&\quad = \|v + \Phi_{\tau,\sigma}^{-1} z - \left(v' + \Phi_{\tau,\sigma}^{-1} z'\right)\|_{\Phi_{\tau,\sigma}} \|v - v'\|_{\Phi_{\tau,\sigma}} \\
&\quad \overset{(i)}{\geq} \left\langle v + \Phi_{\tau,\sigma}^{-1} z - \left(v' + \Phi_{\tau,\sigma}^{-1} z'\right), v - v' \right\rangle_{\Phi_{\tau,\sigma}} \\
&\quad \overset{(ii)}{\geq} \|v - v'\|^2_{\Phi_{\tau,\sigma}} + \mu_f \|v_x - v'_x\|^2 + \mu_g \|v_y - v'_y\|^2 \\
&\quad = (v - v')^\top \left(\Phi_{\tau,\sigma} + \begin{bmatrix} \mu_f I & 0 \\ 0 & \mu_g I \end{bmatrix} \right) (v - v') \\
&\quad \overset{(iii)}{\geq} (v - v')^\top \left((1 + \tilde{\rho}) \Phi_{\tau,\sigma} \right) (v - v') \\
&\quad = (1 + \tilde{\rho}) \|v - v'\|^2_{\Phi_{\tau,\sigma}}
\end{aligned}$$

where (i) follows from Cauchy-Schwarz, (ii) from the definition of the weighted inner product and (iii) from (25). Dividing both sides by $(1 + \tilde{\rho})\|v - v'\|$, we conclude that $\mathcal{B}$ is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with constant ρ (i.e., A1 in Proposition 6, holds with $\Psi_{\mathrm{b}} = \mathbf{0}$, $\gamma = 0$, $\rho_b = \rho$). Then the contractivity result follows by Proposition 6 (we recall that A2 in Proposition 6 holds with $\Psi_{\mathrm{f}} = \mathbf{0}$, $\rho_{\mathrm{f}} = 1$).

Concerning the optimal step sizes, first fix $\tau\sigma\|A\|^2(1 + \alpha)^2 = 1$, for some $\alpha \geq \epsilon$. We now want to maximize κ by choosing opportunely σ and τ, where we recall that

$$\kappa = \left(\frac{\mu_f \tau + \mu_g \sigma - \sqrt{(\mu_f \tau - \mu_g \sigma)^2 + 4\|A\|^2 \mu_f \mu_g \tau^2 \sigma^2}}{2(1 - \sigma\tau\|A\|^2)} \right) > 0.$$

Because $\sigma\tau = \frac{1}{\|A\|^2(1+\alpha)^2}$ is a constant independent of τ and σ, maximizing κ is equivalent to maximizing

$$(\mu_f \tau + \mu_g \sigma) - \sqrt{(\mu_f \tau + \mu_g \sigma)^2 - K},$$

where K is a constant independent of σ and τ. In turn, the latter expression is decreasing in $(\mu_f \tau + \mu_g \sigma)$. Minimizing $\mu_f \tau + \mu_g \sigma = \mu_f \tau + \frac{\mu_g}{\|A\|^2(1+\alpha)^2 \tau}$ is a convex optimization problem in τ, whose solution gives the step size $\tau = \frac{1}{(1+\alpha)\|A\|}\sqrt{\frac{\mu_g}{\mu_f}}$, and thus $\sigma = \frac{1}{(1+\alpha)\|A\|}\sqrt{\frac{\mu_f}{\mu_g}}$. For this values of the step sizes, we get

$$\kappa = \frac{2\frac{\sqrt{\mu_f \mu_g}}{(1+\alpha)\|A\|} - \sqrt{\frac{4\mu_f \mu_g}{\|A\|^2(1+\alpha)^4}}}{2\left(1 - \frac{1}{(1+\alpha)^2}\right)} = \frac{\sqrt{\mu_f \mu_g}}{(\alpha + 2)\|A\|} < \frac{\sqrt{\mu_f \mu_g}}{(\alpha + 1)\|A\|} = \mu_f \tau = \mu_f g. \tag{26}$$

We conclude that the choice $\tau = \frac{1}{(1+\alpha)\|A\|}\sqrt{\frac{\mu_g}{\mu_f}}$, $\sigma = \frac{1}{(1+\alpha)\|A\|}\sqrt{\frac{\mu_f}{\mu_g}}$ maximizes ρ, provided that $\tau\sigma\|A\|^2(1+\alpha)^2 = 1$. Because the optimal rate is increasing in α (i.e., the optimal κ in (26) is decreasing in α), we conclude that the optimal choice is $\alpha = \epsilon$, which is the desired result.

C.5.2 Item (ii): Algorithm 1 under C2

For all $\omega, \omega' \in \mathbb{R}^{n+m}$ and $v \in F_{\mathrm{b}}(\omega)$, $v' \in F_{\mathrm{b}}(\omega')$ we have

$$\begin{aligned}
\|\Phi_{\tau,\sigma}^{-1} v - \Phi_{\tau,\sigma}^{-1} v'\|^2_{\Phi_{\tau,\sigma}} &= \left\langle \Phi_{\tau,\sigma}^{-1}(v - v'), v - v' \right\rangle \\
&\geq \lambda_{\min}\left(\Phi_{\tau,\sigma}^{-1}\right) \|v - v'\|^2 \\
&\overset{(i)}{\geq} \frac{1}{\|\Phi_{\tau,\sigma}\| R_2^2} \|\omega - \omega'\|^2
\end{aligned}$$

$$\geq \frac{1}{\|\Phi_{\tau,\sigma}\|^2 R_2^2}\|\omega-\omega'\|^2_{\Phi_{\tau,\sigma}},$$

where in (i) Lemma 1 was used. The operator $\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}$ is therefore inverse Lipschitz with constant $S_2 = \frac{1}{\|\Phi_{\tau,\sigma}\|R_2}$ with respect to $\|\cdot\|_{\Phi_{\tau,\sigma}}$. Furthermore, since F_b is monotone (in $\langle\cdot,\cdot\rangle$), then $\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}$ is monotone in $\langle\cdot,\cdot\rangle_{\Phi_{\tau,\sigma}}$ by Lemma 9. Therefore we can repeat the proof of Proposition 1, with the only caution of replacing the unweighted norm and inner product with $\|\cdot\|_{\Phi_{\tau,\sigma}}$ and $\langle\cdot,\cdot\rangle_{\Phi_{\tau,\sigma}}$, to conclude that $\mathcal{B} = (\mathrm{Id}+\Phi_{\tau,\sigma}^{-1}F\mathrm{b})$ is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with constant

$$\frac{\|\Phi_{\tau,\sigma}\|R_2}{\sqrt{(\|\Phi_{\tau,\sigma}\|R_2)^2+1}} \leq \frac{\zeta_{\tau,\sigma}R_2}{\sqrt{(\zeta_{\tau,\sigma}R_2)^2+1}} = \rho, \tag{27}$$

where we used the bound $\|\Phi_{\tau,\sigma}\| \leq \zeta_{\tau,\sigma}$ in Proposition 3. In other terms, A1 in Proposition 6 holds with $\Psi_\mathrm{b} = \mathbf{0}$, $\gamma = 0$, $\rho_b = \rho$. Then, the contractivity result follows by Proposition 6 (we recall that A2 in Proposition 6 holds with $\Psi_\mathrm{f} = \mathbf{0}$, $\rho_\mathrm{f} = 1$).

To find the best step sizes, we note that ρ is decreasing in $\zeta_{\tau,\sigma}$. Thus, the best step sizes are found by minimizing $\zeta_{\tau,\sigma}$ in (18), i.e., maximizing $\min\{\tau,\sigma\}$ subject to $\tau\sigma\|A\|^2(1+\epsilon)^2 \leq 1$.

C.5.3 Item (iii): Algorithm 1 under C3

Identical to the proof in Appendix C.5.2, with the only caution of replacing the constant R_2 from Lemma 1 with R_3 from Lemma 2. □

C.6 Proof of Lemma 3

Let $\tau\sigma\|A\|^2 < 1$ and, for any $\omega,\omega' \in \mathbb{R}^{n+m}$, define $u = \mathrm{J}_{\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}}(\omega)$ and $u' = \mathrm{J}_{\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}}(\omega')$ which implies $u + \Phi_{\tau,\sigma}^{-1}F_\mathrm{b}(u) \ni \omega$ and $u' + \Phi_{\tau,\sigma}^{-1}F_\mathrm{b}(u') \ni \omega'$. Let $u = (u_x, u_y)$ and $u' = (u'_x, u'_y)$. Then, for some $v_x \in \partial f(u_x)$ and $v'_x \in \partial f(u'_x)$, we get

$$\begin{aligned}
&\|\omega-\omega'\|^2_{\Phi_{\tau,\sigma}}\\
&= \|u+\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}(u)-u'-\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}(u')\|^2_{\Phi_{\tau,\sigma}}\\
&= \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \|\Phi_{\tau,\sigma}^{-1}(F_\mathrm{b}(u)-F_\mathrm{b}(u'))\|^2_{\Phi_{\tau,\sigma}} + 2\left\langle u-u', \Phi_{\tau,\sigma}^{-1}(F_\mathrm{b}(u)-F_\mathrm{b}(u'))\right\rangle_{\Phi_{\tau,\sigma}}\\
&= \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \left\langle \Phi_{\tau,\sigma}^{-1}(F_\mathrm{b}(u)-F_\mathrm{b}(u')), F_\mathrm{b}(u)-F_\mathrm{b}(u')\right\rangle + 2\left\langle u-u', F_\mathrm{b}(u)-F_\mathrm{b}(u')\right\rangle\\
&\overset{(i)}{\geq} \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \lambda_{\min}\left(\Phi_{\tau,\sigma}^{-1}\right)\|\begin{bmatrix} v_x + A^\top u_y - v'_x - A^\top u'_y \\ -A(u_x-u'_x)\end{bmatrix}\|^2\\
&\quad +2\left\langle u_x-u'_x, v_x-v'_x\right\rangle + \cancel{2\left\langle u_x-u'_x, A^\top(u_y-u'_y)\right\rangle} - \cancel{2\left\langle u_y-u'_y, A(u_x-u'_x)\right\rangle}\\
&\geq \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \lambda_{\min}\left(\Phi_{\tau,\sigma}^{-1}\right)\|A(u_x-u'_x)\|^2 + 2\left\langle u_x-u'_x, v_x-v'_x\right\rangle\\
&\overset{(ii)}{\geq} \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \lambda_{\min}\left(\Phi_{\tau,\sigma}^{-1}\right)\lambda_{\min}\left(A^\top A\right)\|u_x-u'_x\|^2 + 2\left\langle u_x-u'_x, v_x-v'_x\right\rangle\\
&\overset{(iii)}{\geq} \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \lambda_{\min}\left(\Phi_{\tau,\sigma}^{-1}\right)\lambda_{\min}\left(A^\top A\right)\|u_x-u'_x\|^2 + 2\mu_f\|u_x-u'_x\|^2\\
&= \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \left(\lambda_{\min}\left(\Phi_{\tau,\sigma}^{-1}\right)\lambda_{\min}\left(A^\top A\right)+2\mu_f\right)\|\begin{bmatrix} u_x-u'_x \\ 0\end{bmatrix}\|^2\\
&\overset{(iv)}{=} \|u-u'\|^2_{\Phi_{\tau,\sigma}} + \|u-v\|^2_{\Psi_b}\\
&= \|\mathrm{J}_{\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}}(\omega)-\mathrm{J}_{\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}}(\omega')\|^2_{\Phi_{\tau,\sigma}} + \|\mathrm{J}_{\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}}(\omega)-\mathrm{J}_{\Phi_{\tau,\sigma}^{-1}F_\mathrm{b}}(\omega')\|^2_{\Psi_b},
\end{aligned}$$

where in (i) the fact that $\Phi_{\tau,\sigma}^{-1}$ is invertible was used, (ii) holds since the smallest singular value is non-negative, (iii) follows by (strong) convexity of f and (iv) follows from the definition of Ψ_b. □

C.7 Proof of Lemma 4

We need the following additional result.

Lemma 10. *Let $\xi_1 = \frac{1}{\sigma} - \tau\|A\|^2$. Then, for all $\omega = (x, y), \omega' = (x', y') \in \mathbb{R}^{n+m}$,*

$$\langle F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')), \omega - \omega' \rangle \geq \frac{L_g \mu_g}{L_g + \mu_g} \| \begin{bmatrix} 0 \\ y - y' \end{bmatrix} \|^2 + \frac{\xi_1}{L_g + \mu_g} \|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\|^2_{\Phi^{-1}_{\tau,\sigma}}.$$

Proof. Let $h(\omega) = g(y) - \frac{\mu_g}{2}\|y\|^2$. Since h is $(L_g - \mu_g)$-smooth, we have the cocoercivity bound (Bauschke and Combettes, 2017, Th. 18.15(v))

$$\langle \nabla g(y) - \nabla g(y'), y - y' \rangle = \langle F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega'), \omega - \omega' \rangle \geq \frac{1}{L_g - \mu_g} \|\nabla h(\omega) - \nabla h(\omega')\|^2.$$

Rearranging the terms gives

$$\langle F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')), \omega - \omega' \rangle \geq \frac{L_g \mu_g}{L_g + \mu_g} \| \begin{bmatrix} 0 \\ y - y' \end{bmatrix} \|^2 + \frac{1}{L_g + \mu_g} \|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\|^2.$$

The conclusion follows because $F_{\mathrm{f}} = (\mathbf{0}, \nabla g)$ and, by computing the block inverse of Φ, the right lower block of $\Phi^{-1}_{\tau,\sigma}$ is $(\frac{1}{\sigma} I - \tau AA^\top)$, so that

$$\|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\| \leq \xi_1 \|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\|_{\Phi^{-1}_{\tau,\sigma}}.$$

□

We are ready to prove Lemma 4.

Proof. For any $\omega, \omega' \in \mathbb{R}^{n+m}$, with $\omega = \begin{bmatrix} x \\ y \end{bmatrix}$ and $\omega' = \begin{bmatrix} x' \\ y' \end{bmatrix}$, we have

$$\begin{aligned} &\| \left(I - \Phi^{-1}_{\tau,\sigma} F_{\mathrm{f}}\right)(\omega) - \left(I - \Phi^{-1}_{\tau,\sigma} F_{\mathrm{f}}\right)(\omega')\|^2_{\Phi_{\tau,\sigma}} \\ &= \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}} + \|\Phi^{-1}_{\tau,\sigma}(F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega'))\|^2_{\Phi_{\tau,\sigma}} - 2\left\langle \omega - \omega', F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega') \right\rangle \\ &\leq \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}} - \frac{2L_g\mu_g}{L_g + \mu_g} \| \begin{bmatrix} 0 \\ y - y' \end{bmatrix} \|^2 + \left(1 - \frac{2\xi_1}{L_g + \mu_g}\right) \|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\|_{\Phi^{-1}_{\tau,\sigma}}, \end{aligned}$$

where in the last line Lemma 10 and the fact that $\|\Phi^{-1}_{\tau,\sigma}(F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega'))\|^2_{\Phi_{\tau,\sigma}} = \|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\|^2_{\Phi^{-1}_{\tau,\sigma}}$ was used. We have to distinguish between two cases:

- $1 - \frac{2\xi_1}{L_g + \mu_g} \leq 0$:

  In this case we can use the bound $\|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\|^2_{\Phi^{-1}_{\tau,\sigma}} \geq \frac{\mu_g^2}{\xi_2} \| \begin{bmatrix} 0 \\ y - y' \end{bmatrix} \|^2$, with $\xi_2 = \frac{1}{\sigma} I - \tau \mu_A \geq \lambda_{\max}(\frac{1}{\sigma} I - \tau AA^\top)$, which follows again by definition of F_{f} and block inverse of $\Phi_{\tau,\sigma}$. This implies

$$\begin{aligned} &\| \left(I - \Phi^{-1}_{\tau,\sigma} F_{\mathrm{f}}\right)(\omega) - \left(I - \Phi^{-1}_{\tau,\sigma} F_{\mathrm{f}}\right)(\omega')\|^2_{\Phi_{\tau,\sigma}} \\ &\leq \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}} - \left(\frac{2L_g\mu_g}{L_g + \mu_g} + \frac{2\mu_g^2 \xi_1}{\xi_2(L_g + \mu_g)} - \frac{\mu_g^2}{\xi_2}\right) \| \begin{bmatrix} 0 \\ y - y' \end{bmatrix} \|^2. \end{aligned}$$

- $1 - \frac{2\xi_1}{L_g + \mu_g} > 0$:

  Using $\|F_{\mathrm{f}}(\omega) - F_{\mathrm{f}}(\omega')\|^2_{\Phi^{-1}_{\tau,\sigma}} \leq \frac{L_g^2}{\xi_1} \| \begin{bmatrix} 0 \\ y - y' \end{bmatrix} \|^2$ we get

$$\begin{aligned} &\| \left(I - \Phi^{-1}_{\tau,\sigma} F_{\mathrm{f}}\right)(\omega) - \left(I - \Phi^{-1}_{\tau,\sigma} F_{\mathrm{f}}\right)(\omega')\|^2_{\Phi_{\tau,\sigma}} \\ &\leq \|\omega - \omega'\|^2_{\Phi_{\tau,\sigma}} - \left(\frac{2L_g\mu_g}{L_g + \mu_g} + \frac{2L_g^2}{L_g + \mu_g} - \frac{L_g^2}{\xi_1}\right) \| \begin{bmatrix} 0 \\ y - y' \end{bmatrix} \|^2. \end{aligned} \tag{28}$$

The last addend is negative only if $\xi_1 > \frac{L_g}{2}$, or equivalently

$$\sigma < \frac{2}{2\tau\|A\|^2 + L_g}.$$

Therefore, if the latter bounds on the step sizes holds, the forward step fulfills

$$\| \left(I - \Phi_{\tau,\sigma}^{-1} F_{\mathrm{f}}\right)(\omega) - \left(I - \Phi_{\tau,\sigma}^{-1} F_{\mathrm{f}}\right)(\omega')\|_{\Phi_{\tau,\sigma}}^2 \leq \|\omega - \omega'\|_{\Phi_{\tau,\sigma}}^2 - \|\omega - \omega'\|_{\Psi_{\mathrm{f}}}^2,$$

where $\Psi_{\mathrm{f}} = \operatorname{diag}(\mathbf{0}, \gamma_y I_m)$ and

$$0 < \gamma_y = \begin{cases} \frac{2L_g\mu_g}{L_g+\mu_g} + \frac{\mu_g^2(2\xi_1 - L_g - \mu_g)}{\xi_2(L_g+\mu_g)} & \text{if } \ \xi_1 \geq \frac{L_g+\mu_g}{2} \\ \frac{2L_g\mu_g}{L_g+\mu_g} - \frac{L_g^2(L_g+\mu_g-2\xi_1)}{(L_g+\mu_g)\xi_1} & \text{if } \ \xi_1 < \frac{L_g+\mu_g}{2}, \end{cases}$$

which concludes the proof. □

C.8 Proof of Theorem 3

As in the proof of Theorem 2, for any $\tau, \sigma > 0$, the fixed points of Algorithm 2 coincide with the saddle-points of (1). Furthermore, whenever $\sigma\tau\|A\|^2 < 1$ (which is ensured by the assumptions on the step sizes in Theorem 3), $\Phi_{\tau,\sigma} \succ 0$ and Algorithm 2 can be recast as the forward-backward iteration in (11), with $F_{\mathrm{f}}(\omega) = (\mathbf{0}, \nabla g(y))$, $F_{\mathrm{b}}(\omega) = (\nabla f(x) + A^\top y, -Ax)$.

C.8.1 Item(i): Algorithm 2 under C1 or C2

Under either C1 or C2, $\mathcal{B}$ satisfies A1 in Proposition 2 with Φ_{b} as in Lemma 3, and $\mathcal{F}$ satisfies A2 in Proposition 2 with Φ_{f} as in Lemma 4. Furthermore,

$$\begin{aligned} \Psi_{\mathrm{b}} + \Psi_{\mathrm{f}} = \operatorname{diag}(\gamma_x I_n, \gamma_y I_m) &\geq \frac{\min\{\gamma_x, \gamma_y\}}{\|\Phi_{\tau,\sigma}\| + \gamma_x} \Psi_{\tau,\gamma} \\ &\geq \frac{\min\{\gamma_x, \gamma_y\}}{\zeta_{\tau,\sigma} + \gamma_x} (\Phi_{\tau,\sigma} + \Psi_{\mathrm{b}}), \end{aligned}$$

and the conclusion readily follows by Proposition 2.

C.8.2 Item (ii): Algorithm 2 under C3

First, we show that the backward step $\mathrm{J}_{\Phi_{\tau,\sigma}^{-1} F_{\mathrm{b}}}$ is contractive. The proof is analogous to Lemma 2 and Theorem 2(iii). Define the following matrix $\Psi_\varepsilon \in \mathbb{R}^{2n \times 2n}$, for $\varepsilon > 0$ small enough to be chosen,

$$\Psi_\varepsilon = \begin{bmatrix} I_n & \varepsilon A^\top \\ -\varepsilon A & I_n \end{bmatrix}.$$

Then we have, for $\omega = (x, y)$, $\omega' = (x', y') \in \mathbb{R}^{2n}$,

$$\begin{aligned} &\|F_{\mathrm{b}}(\omega) - F_{\mathrm{b}}(\omega')\|\|\Psi_\varepsilon\|\|\omega - \omega'\| \\ &\geq \ \| \begin{bmatrix} \nabla f(x) - \nabla f(x') + A^\top(y - y') \\ -A(x - x') \end{bmatrix} \|\| \begin{bmatrix} x - x' + \varepsilon A^\top(y - y') \\ y - y' - \varepsilon A(x - x') \end{bmatrix} \| \\ &\overset{(i)}{\geq} \left\langle \begin{bmatrix} \nabla f(x) - \nabla f(x') + A^\top(y - y') \\ -A(x - x') \end{bmatrix}, \begin{bmatrix} x - x' + \varepsilon A^\top(y - y') \\ y - y' - \varepsilon A(x - x') \end{bmatrix} \right\rangle \\ &= \ \langle \nabla f(x) - \nabla f(x'), x - x' \rangle + \varepsilon \left\langle \nabla f(x) - \nabla f(x'), A^\top(y - y') \right\rangle + \varepsilon\|A^\top(y - y')\|^2 + \varepsilon\|A(x - x')\|^2 \\ &\overset{(ii)}{\geq} \frac{1}{L_f}\|\nabla f(x) - \nabla f(x')\|^2 + \varepsilon \left\langle \nabla f(x) - \nabla f(x'), A^\top(y - y') \right\rangle + \varepsilon\|A^\top(y - y')\|^2 + \varepsilon\|A(x - x')\|^2 \\ &\overset{(iii)}{\geq} \left(\frac{1}{L_f} - \frac{\|A\|\varepsilon\delta}{2}\right)\|\nabla f(x) - \nabla f(x')\|^2 + \varepsilon\left(\lambda_{\min}\left(AA^\top\right) - \frac{\|A\|}{2\delta}\right)\|y - y'\|^2 \end{aligned}$$

$$
\begin{aligned}
&+\varepsilon\left(\lambda_{\min}\left(A^\top A\right)-\frac{\|A\|}{2\delta}\right)\|x-x'\|^2 \\
&\overset{(iv)}{\geq} C(\varepsilon,\delta)\|\begin{bmatrix} x-x' \\ y-y' \end{bmatrix}\|^2,
\end{aligned}
$$

where (i) follows from Cauchy-Schwarz, (ii) from cocoercivity of ∇f, (iii) from Young's inequality and (iv) holds if $\delta > \frac{\|A\|\mu_A}{2}$, $\varepsilon < \frac{2}{L_f\|A\|\delta}$ and

$$
C(\varepsilon,\delta)=\varepsilon\left(\mu_A-\frac{\|A\|}{2\delta}\right)>0.
$$

Then the proof follows as for Appendix C.5.2. In particular, using $\|\Psi_\varepsilon\| \leq 1+\varepsilon\|A\|$ we have that F_{b} is inverse Lipschitz with constant $\frac{1}{R_3'}$, $R_3' := \frac{1+\varepsilon\|A\|}{C(\varepsilon,\delta)}$, thus $\Phi_{\tau,\sigma}^{-1}F_{\mathrm{b}}$ is inverse Lipschitz in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with constant $\frac{R_3'}{\|\Phi_{\tau,\sigma}\|}$. Since $\Phi_{\tau,\sigma}^{-1}F_{\mathrm{b}}$ is maximally monotone in $\|\cdot\|_{\Phi_{\tau,\sigma}}$, the backward step $\mathcal{B}$ is contractive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ with parameter $\frac{\zeta_{\tau,\sigma}R_3'}{\sqrt{(R_3'\zeta_{\tau,\sigma})^2+1}}$ (see Proposition 1(ii)), namely A1 in Proposition 6 holds with $\rho_{\mathrm{b}}=\rho$, $\Phi_{\mathrm{b}}=\mathbf{0}$.

It is left to show, that the forward step $\mathcal{F}=I-\Phi_{\tau,\sigma}^{-1}F_{\mathrm{f}}$ is nonexpansive, namely that A2 in Proposition 6 holds with $\rho_{\mathrm{f}}=1$, $\Psi_{\mathrm{f}}=\mathbf{0}$. If $L_g=0$, this holds trivially (we recall that $F_{\mathrm{f}}=(\mathbf{0},\nabla g)$). If $L_g\neq 0$, then this holds whenever $\sigma<\frac{2}{L_g+2\|A\|}$ by (28). Then, the contractivity result in Theorem 3(ii) follows by Proposition 2.

To find the best step sizes, we note that ρ is decreasing in $\zeta_{\tau,\sigma}$. Thus, the best step sizes are found by minimizing $\zeta_{\tau,\sigma}$ in (18), i.e., maximizing $\min\{\tau,\sigma\}$ subject to $\tau\sigma\|A\|^2+\frac{L_g}{2}\sigma\leq 1$ and $\tau\sigma\|A\|^2(1+\epsilon)^2\leq 1$. Clearly $\sigma=\tau$ at the solution. Hence the best step size $\tau=\sigma$ is obtained as the maximum value that satisfies $\sigma^2\|A\|^2+\frac{L_g}{2}\sigma\leq 1$ and $\sigma^2\|A\|^2(1+\epsilon)^2\leq 1$, or equivalently $\sigma\leq\frac{\sqrt{L_g^2+16\|A\|^2}-L_g}{4\|A\|^2}$ and $\sigma\leq\frac{1}{(1+\epsilon)\|A\|}$, which is the claim. $\square$

C.9 Proof of Lemma 5

First, let us assume $L_f, L_g\neq 0$. Let us recall that, as in the proof of Lemma 4, for any $\omega=(x,y),\omega'=(x',y')\in\mathbb{R}^{n+m}$,

$$
\begin{aligned}
&\langle\omega-\omega',F(\omega)-F(\omega')\rangle \\
\geq\ &\frac{L_f\mu_f}{L_f+\mu_f}\|x-x'\|^2+\frac{1}{L_f+\mu_f}\|\nabla f(x)-\nabla f(x')\|^2+\frac{L_g\mu_g}{L_g+\mu_g}\|y-y'\|^2+\frac{1}{L_g+\mu_g}\|\nabla g(y)-\nabla g(y')\|^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\|\mathcal{F}(\omega)-\mathcal{F}(\omega')\|_{\Phi_{\tau,\sigma}} \\
&\leq\|\omega-\omega'\|_{\Phi_{\tau,\sigma}}-2\frac{L_f\mu_f}{L_f+\mu_f}\|x-x'\|^2-2\frac{L_g\mu_g}{L_g+\mu_g}\|y-y'\|^2-2\frac{1}{L_f+\mu_f}\|\nabla f(x)-\nabla f(x')\|^2 \\
&\quad-2\frac{1}{L_g+\mu_g}\|\nabla g(y)-\nabla g(y')\|^2+\|F_{\mathrm{f}}(\omega)-F_{\mathrm{f}}(\omega')\|^2_{\Phi_{\tau,\sigma}^{-1}}.
\end{aligned}
$$

To bound the last three addends, we look for the smallest (possibly negative) constant α_x, α_y , such that

$$
\Phi_{\tau,\sigma}^{-1}-\operatorname{diag}\left(2\tfrac{1}{L_f+\mu_f}I_n,2\tfrac{1}{L_g+\mu_g}I_m\right)\preccurlyeq\operatorname{diag}(\alpha_x I_n,\alpha_y I_m) \tag{29}
$$

(since $\Phi_{\tau,\sigma}^{-1}\succ 0$, this implies $2\frac{1}{L_f+\mu_f}+\alpha_x>0$, $2\frac{1}{L_g+\mu_g}+\alpha_y>0$). By double application of the Schur complement, the above condition is equivalent to

$$
\Phi_{\tau,\sigma}-\operatorname{diag}\left(\tfrac{L_f+\mu_f}{2+\alpha_x(L_f+\mu_f)}I_n,\tfrac{L_g+\mu_g}{2+\alpha_y(L_g+\mu_g)}I_m\right)\succcurlyeq 0. \tag{30}
$$

By computing the block-matrix determinant in (30), we obtain that a sufficient condition for this to hold is that, for any $\nu>0$,

$$
\frac{1}{\tau}-\tfrac{L_f+\mu_f}{2+\alpha_x(L_f+\mu_f)}\geq\|A\|\nu
$$

$$\frac{1}{\sigma} - \tfrac{L_g+\mu_g}{2+\alpha_y(L_g+\mu_g)} \geq \|A\|\frac{1}{\nu}.$$

Solving for α_x and α_y (noting that $1-\tau\|A\|>0$), we obtain the smallest values

$$\alpha_x = \frac{\tau}{1-\tau\|A\|\nu} - \frac{2}{L_f+\mu_f}, \qquad \alpha_y = \frac{\sigma}{1-\sigma\|A\|\frac{1}{\nu}} - \frac{2}{L_g+\mu_g},$$

and get

$$\begin{aligned}&\|\mathcal{F}(\omega)-\mathcal{F}(\omega')\|_{\Phi_{\tau,\sigma}}\\ &\leq \|\omega-\omega'\|_{\Phi_{\tau,\sigma}} - 2\frac{L_f\mu_f}{L_f+\mu_f}\|x-x'\|^2 - 2\frac{L_g\mu_g}{L_g+\mu_g}\|y-y'\|^2 + \alpha_x\|\nabla f(x)-\nabla f(x')\|^2 + \alpha_y\|\nabla g(y)-\nabla g(y')\|^2.\end{aligned}$$

Note that α_x, α_y might be positive or negative, depending on the step sizes τ, σ. In particular, if $\tau \leq \frac{2}{L_f+\mu_f+2\|A\|\nu}$, then α_x is nonpositive, and we can use the bound $\|\nabla f(x)-\nabla f(x')\| \geq \mu_f\|x-x'\|$, derived from strong convexity of f (with μ_f possibly zero); and if $\tau > \frac{2}{L_f+\mu_f+2\|A\|\nu}$, then α_x is positive, and we use the bound $\|\nabla f(x)-\nabla f(x')\| \leq L_f\|x-x'\|$, and similarly for α_y. Finally, we get

$$\|\mathcal{F}(\omega)-\mathcal{F}(\omega')\|_{\Phi_{\tau,\sigma}} \leq \|\omega-\omega'\|_{\Phi_{\tau,\sigma}} - \beta_x\|x-x'\|^2 - \beta_y\|y-y'\|^2, \tag{31}$$

where

$$\beta_x = \begin{cases} 2\frac{L_f\mu_f}{L_f+\mu_f} - \alpha_x\mu_f^2 & \text{if } \tau \leq \frac{2}{L_f+\mu_f+2\|A\|\nu} \\ 2\frac{L_f\mu_f}{L_f+\mu_f} - \alpha_x L_f^2 & \text{if } \tau > \frac{2}{L_f+\mu_f+2\|A\|\nu} \end{cases}$$

$$\beta_y = \begin{cases} 2\frac{L_g\mu_g}{L_g+\mu_g} - \alpha_y\mu_g^2 & \text{if } \sigma \leq \frac{2}{L_g+\mu_g+2\|A\|\frac{1}{\nu}} \\ 2\frac{L_g\mu_g}{L_g+\mu_g} - \alpha_y L_g^2 & \text{if } \sigma > \frac{2}{L_g+\mu_g+2\|A\|\frac{1}{\nu}}, \end{cases}$$

and it is easily verified that $\beta_x,\beta_y>0$ whenever $\mu_f,\mu_g>0$ and $\tau<\frac{2}{L_f+2\|A\|\nu}$, $\sigma<\frac{2}{L_g+2\|A\|\frac{1}{\nu}}$ (and $\beta_x,\beta_y=0$ if $\tau=\frac{2}{L_f+2\|A\|\nu}$, $\sigma=\frac{2}{L_g+2\|A\|\frac{1}{\nu}}$).

Finally, let us note that the statement still holds with $\beta_x=0$ whenever $L_f=0$, and with $\beta_y=0$ whenever $L_g=0$ (repeat the same proof by omitting the terms corresponding to the zero Lipschitz constants). □

C.10 Proof of Theorem 4

For any $\tau,\sigma>0$, the fixed points of Algorithm 3 coincide with the saddle points of (1), as in the proof of Theorem 2. Furthermore, whenever $\sigma\tau\|A\|^2<1$ (which is ensured by the assumptions on the step sizes in Theorem 4), $\Phi_{\tau,\sigma}\succ 0$ and Algorithm 3 can be recast as the forward-backward iteration in (11), with $F_{\mathrm{f}}=(\nabla f,\nabla g)$, $F_{\mathrm{b}}=(A^\top y,-Ax)$.

C.10.1 Item(i): Algorithm 3 under C1

Under C1, the forward step $\mathcal{F}$ is contractive, as a direct consequence of Lemma 5. In particular, its contractivity rate $\rho>0$ in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ can be found by solving

$$\Phi_{\tau,\sigma} - \operatorname{diag}(\beta_x I_n, \beta_y I_m) \preccurlyeq \rho\Phi,$$

which, via a Schur complement argument, can be relaxed as

$$\beta_x - \frac{1}{\tau}(1-\rho) \geq (1-\rho)\|A\|\nu, \qquad \beta_y - (1-\rho)\frac{1}{\sigma} \geq (1-\rho)\|A\|\tfrac{1}{\nu},$$

equivalently

$$\rho \geq 1 - \min\left\{\frac{\beta_x\tau}{1+\tau\|A\|\nu}, \frac{\beta_y\sigma}{1+\sigma\|A\|\frac{1}{\nu}}\right\}.$$

Hence, A2 in Proposition 6 holds with $\Psi_{\mathrm{f}} = \mathbf{0}$ and $\rho_{\mathrm{f}} = \rho$.

It is only left to show that the backward step $\mathcal{B}$ is nonexpansive, namely that A1 in Proposition 6 holds with $\Psi_{\mathrm{b}} = \mathbf{0}$ and $\rho_{\mathrm{b}} = 1$. This is a consequence of the fact that the operator $\Phi_{\tau,\sigma}^{-1} F_{\mathrm{b}}$ is monotone in $\langle\cdot,\cdot\rangle_{\Phi}$ (since F_{b} is monotone in $\langle\cdot,\cdot\rangle$ and by Lemma 9), and hence its resolvent $\mathcal{B}$ is nonexpansive in $\|\cdot\|_{\Phi_{\tau,\sigma}}$ (Bauschke and Combettes, 2017, Prop. 23.8). Then the contractivity result follows by Proposition 6.

C.10.2 Item(ii): Algorithm 3 under C2

By specializing Lemma 3 to the case $f = 0$, we obtain that A1 in Proposition 2 holds with $\Psi_{\mathrm{b}} = \operatorname{diag}\left(\frac{\mu_A}{\zeta_{\tau,\sigma}} I_n, \mathbf{0}\right)$.

On the other hand, A2 holds with Ψ_{f} as in Lemma 5 (where $\beta_x = 0$ in general).

Finally,

$$\begin{aligned}\Psi_{\mathrm{b}} + \Psi_{\mathrm{f}} = \operatorname{diag}\left(\frac{\mu_A}{\zeta_{\tau,\sigma}} I_n, \beta_y I_m\right) &\succeq \frac{\min\left\{\frac{\mu_A}{\zeta_{\tau,\sigma}}, \beta_y\right\}}{\|\Psi_{\mathrm{b}} + \Phi_{\tau,\sigma}\|}(\Psi_{\mathrm{b}} + \Phi_{\tau,\sigma}) \\ &\succeq \frac{\min\left\{\frac{\mu_A}{\zeta_{\tau,\sigma}}, \beta_y\right\}}{\zeta_{\tau,\sigma} + \mu_A}(\Psi_{\mathrm{b}} + \Phi_{\tau,\sigma}),\end{aligned}$$

and the conclusion readily follows by Proposition 2.

C.10.3 Item(iii): Algorithm 3 under C3

We start by an auxiliary result, which refines Lemma 1.

Lemma 11. *Let C3 hold. Then, the operator $F_b(\omega) = (A^\top y, -Ax)$ is $\sqrt{\mu_A}$-inverse Lipschitz.*

Note that $\Phi_{\tau,\sigma}^{-1} F_{\mathrm{b}}$ is monotone in $\langle\cdot,\cdot\rangle_{\Phi_{\tau,\sigma}}$ by Lemma 9. Furthermore, F_{b} is μ_A-inverse Lipschitz according to Lemma 11, and hence $\Phi_{\tau,\sigma}^{-1} F_{\mathrm{b}}$ is inverse Lipschitz in $\langle\cdot,\cdot\rangle_{\Phi_{\tau,\sigma}}$ with constant $\frac{\sqrt{\mu_A}}{\|\Phi_{\tau,\sigma}\|}$. Hence, the backward step $\mathrm{J}_{\Phi_{\tau,\sigma}^{-1} F_{\mathrm{b}}}$ is contractive in $\|\cdot\|_{\Phi}$ with factor $\frac{\|\Phi_{\tau,\sigma}\|}{\sqrt{\|\Phi_{\tau,\sigma}\|^2 + \mu_A}} \leq \frac{\zeta_{\tau,\sigma}}{\sqrt{\zeta_{\tau,\sigma}^2 + \mu_A}}$ (the proof is identical to that of Proposition 1, with the only caution of replacing $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ with $\langle\cdot,\cdot\rangle_{\Phi_{\tau,\sigma}}$ and $\|\cdot\|_{\Phi_{\tau,\sigma}}$). Therefore, A1 in Proposition 6 is satisfied with $\Psi_{\mathrm{b}} = \mathbf{0}$, $\rho_{\mathrm{b}} = \rho$.

On the other hand, by Lemma 5 (with $\beta_x = \beta_y = 0$), A2 in Proposition 6 is satisfied with $\Psi_{\mathrm{f}} = \mathbf{0}$, $\rho_{\mathrm{f}} = 1$ (i.e., the forward step is nonexpansive). Then, the contractivity result follows by Proposition 2.

To find the best step sizes, we note that ρ is decreasing in $\zeta_{\tau,\sigma}$. Thus, the best step sizes are found by minimizing $\zeta_{\tau,\sigma}$ in (18), i.e., maximizing $\min\{\tau, \sigma\}$ subject to $\tau \leq \frac{2}{L_f + +2\|A\|\nu}$, $\sigma \leq \frac{2}{L_g + 2\|A\|\frac{1}{\nu}}$, $\tau\sigma\|A\|^2(1+\epsilon)^2 \leq 1$. Clearly $\sigma = \tau$ at a solution (otherwise, there exist $\tau' = \sigma'$ satisfying the step size bounds for some ν, such that $\min\{\sigma', \tau'\} > \min\{\sigma, \tau\}$). Therefore, the the optimal $\bar{\nu}$ is obtained by solving for ν the equation $\tau = \frac{2}{L_f + 2\|A\|\nu} = \frac{2}{L_g + 2\|A\|\frac{1}{\nu}} = \sigma$.

Note that the bound $\tau\sigma\|A\|^2(1+\epsilon)^2 \leq 1$ is used to ensure $\Phi_{\tau,\sigma} \succ 0$ when both L_f and L_g are zero. For this particular case, similar considerations to Remark 1 hold for the choice of ϵ; otherwise, $\epsilon = 0$ can be chosen (since any ϵ small enough would be inconsequential for the bounds and optimal choice of step sizes). □

D Extensions

D.1 Gradient Descent-Ascent

We assume that C2 holds (namely, that g is L_g-smooth and μ_g-strongly convex, and A has full column rank) and that f is L_f-smooth. Under these assumptions, the R-linear convergence of the PDG method for small-enough step sizes was studied first proved by Du and Hu (2019). Here, we show that the PDG method is actually *contractive* (hence, converges Q-linearly). The proof is very simple: we show that the saddle-point operator in (10) is strongly monotone and Lipschitz continuous in a weighted space. Then the contractivity of the method follows by known results for the forward step, e.g., (Bauschke and Combettes, 2017, Prop. 26.16).

Algorithm 4 Primal-dual gradient method

Require: $x^0 \in \mathbb{R}^n, y^0 \in \mathbb{R}^m$, step size $\alpha > 0$

1: **for** $k = 0, 1, \cdots$ **do**

2: $\quad x^{k+1} = x^k - \alpha \left(\nabla f(x^k) + A^\top y^k\right)$

3: $\quad y^{k+1} = y^k - \alpha \left(\nabla g(y^k) - Ax^k\right)$

4: **end for**

To streamline the presentation, we consider the algorithm in Algorithm 4, which only uses one step size $\alpha > 0$ (instead of two step sizes τ, σ, for which the analysis would be analogous).

Let us define

$$\Phi_\eta = \begin{bmatrix} I_n & -\eta A^\top \\ -\eta A & I_m, \end{bmatrix} \tag{32}$$

where $\eta > 0$ is a parameter to be chosen, and $\Phi_\eta \succ 0$ if $\eta < \frac{1}{\|A\|}$.

Lemma 12. *If $\eta > 0$ is chosen small enough, then $\Phi_\eta \succ 0$ and there is $\mu_\eta > 0$ such that F in* (10) *is μ_η-strongly monotone in the space weighted by Φ_η, i.e., for all $\omega, \omega' \in \mathbb{R}^{n+m}$*

$$\langle F(\omega) - F(\omega'), \omega - \omega' \rangle_{\Phi_\eta} \geq \mu_\eta \|\omega - \omega'\|^2_{\Phi_\eta}.$$

Proof. Let $\omega = (x, y), \omega' = (x', y') \in \mathbb{R}^{n+m}$ and let $\eta < \frac{1}{\|A\|}$, so that $\Phi_\eta \succ 0$. We have

$$\begin{aligned}
&\langle F(\omega) - F(\omega'), \omega - \omega' \rangle_{\Phi_{\tau,\sigma}} \\
=\ & \left\langle \begin{bmatrix} x - x' \\ y - y' \end{bmatrix}, \begin{bmatrix} \nabla f(x) - \nabla f(x') + A^\top (y - y') \\ \nabla g(y) - \nabla g(y') - A(x - x') \end{bmatrix} \right\rangle_{\Phi_\eta} \\
=\ & \left\langle \begin{bmatrix} x - x' - \eta A^\top (y - y') \\ y - y' - \eta A(x - x') \end{bmatrix}, \begin{bmatrix} \nabla f(x) - \nabla f(x') + A^\top (y - y') \\ \nabla g(y) - \nabla g(y') - A(x - x') \end{bmatrix} \right\rangle \\
=\ & \left\langle x - x' - \eta A^\top (y - y'), \nabla f(x) - \nabla f(x') + A^\top (y - y') \right\rangle \\
& + \langle y - y' - \eta A(x - x'), \nabla g(y) - \nabla g(y') - A(x - x') \rangle \\
=\ & \langle x - x', \nabla f(x) - \nabla f(x') \rangle + \langle y - y', \nabla g(y) - \nabla g(y') \rangle \\
& -\eta \left\langle A^\top (y - y'), A^\top (y - y') \right\rangle + \eta \langle A(x - x'), A(x - x') \rangle \\
& -\eta \left\langle A^\top (y - y'), \nabla f(x) - \nabla f(x') \right\rangle - \eta \langle A(x - x'), \nabla g(y) - \nabla g(y') \rangle \\
\overset{(i)}{\geq}\ & \mu_g \|y - y'\|^2 - \eta \left\langle AA^\top (y - y'), y - y' \right\rangle + \eta \left\langle A^\top A(x - x'), x - x' \right\rangle \\
& -\eta \left\langle A^\top (y - y'), \nabla f(x) - \nabla f(x') \right\rangle - \eta \langle A(x - x'), \nabla g(y) - \nabla g(y') \rangle \\
\geq\ & \mu_g \|y - y'\|^2 - \eta \lambda_{\max}(AA^\top) \|y - y'\|^2 + \eta \lambda_{\min}(A^\top A) \|x - x'\|^2 \\
& -\eta \left\langle A^\top (y - y'), \nabla f(x) - \nabla f(x') \right\rangle - \eta \langle A(x - x'), \nabla g(y) - \nabla g(y') \rangle \\
\overset{(ii)}{\geq}\ & \mu_g \|y - y'\|^2 - \eta \lambda_{\max}(AA^\top) \|y - y'\|^2 + \eta \lambda_{\min}(A^\top A) \|x - x'\|^2 \\
& -\eta (L_f + L_g) \|A\| \|x - x'\| \|y - y'\| \\
=\ & \begin{bmatrix} \|x - x'\| \\ \|y - y'\| \end{bmatrix} \begin{bmatrix} \eta \lambda_{\min}(A^\top A) & -\frac{1}{2}\eta (L_f + L_g) \|A\| \\ -\frac{1}{2}\eta (L_f + L_g) \|A\| & \mu_g - \eta \lambda_{\max}(AA^\top) \end{bmatrix} \begin{bmatrix} \|x - x'\| \\ \|y - y'\| \end{bmatrix} \\
\overset{(iii)}{\geq}\ & \lambda_{\min}(M_\eta) \| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \|^2,
\end{aligned}$$

where we used monotononicity of the gradient and strong convexity of g in (i), in (ii) we used Cauchy-Schwarz and strong smoothness of f and g, and (iii) follows by defining

$$M_\eta := \begin{bmatrix} \eta \lambda_{\min}(A^\top A) & -\frac{1}{2}\eta (L_f + L_g) \|A\| \\ -\frac{1}{2}\eta (L_f + L_g) \|A\| & \mu_g - \eta \lambda_{\max}(AA^\top) \end{bmatrix}. \tag{33}$$

We next show that $M_\eta \succ 0$ for small enough η. We have

$$\det(M_\eta) = (\eta \lambda_{\min}(A^\top A))(\mu_g - \eta \lambda_{\max}(AA^\top)) - \frac{1}{4}\eta^2 (L_f + L_g)^2 \|A\|^2,$$

which is positive if and only if

$$\eta < \frac{\mu_g \lambda_{\min}(A^\top A)}{\lambda_{\min}(A^\top A)\lambda_{\max}(AA^\top) + \frac{1}{4}(L_f + L_g)^2\|A\|^2} =: C_M. \tag{34}$$

Since $\eta\lambda_{\min}(A^\top A) > 0$, from Sylvester's criterion it follows that $M_\eta \succ 0$ for all η small enough. In particular, for any $\eta < \min\{\frac{1}{\|A\|}, C_M\}$, the conclusion follows with

$$\mu_\eta = \frac{\lambda_{\min}(M_\eta)}{\|\Phi_\eta\|} > 0. \tag{35}$$

□

Lemma 13. *Let $\eta < \frac{1}{\|A\|}$. Then, the saddle-point operator F in (10) is Lipschitz in $\|\cdot\|_{\Phi_\eta}$ with constant $L_\eta = \sqrt{\frac{\lambda_{\max}(\Phi_\eta)}{\lambda_{\min}(\Phi_\eta)}}\sqrt{\max\{L_f, L_g\}^2 + \|A\|^2} > 0$.*

Proof. Let $x, x' \in \mathbb{R}^n$ and $y, y' \in \mathbb{R}^m$ be arbitrary. We have

$$\begin{aligned}
\frac{1}{\lambda_{\max}(\Phi_\eta)}\|F\left(\begin{bmatrix}x\\y\end{bmatrix}\right) - F\left(\begin{bmatrix}x'\\y'\end{bmatrix}\right)\|^2_{\Phi_\eta} &\le \|F\left(\begin{bmatrix}x\\y\end{bmatrix}\right) - F\left(\begin{bmatrix}x'\\y'\end{bmatrix}\right)\|^2 \\
&= \|\begin{bmatrix}\nabla f(x) - \nabla f(x')\\ \nabla g(y) - \nabla g(y')\end{bmatrix} + \begin{bmatrix}A^\top(y-y')\\ -A(x-x')\end{bmatrix}\|^2 \\
&\le \|\begin{bmatrix}\nabla f(x) - \nabla f(x')\\ \nabla g(y) - \nabla g(y')\end{bmatrix}\|^2 + \|\begin{bmatrix}A^\top(y-y')\\ -A(x-x')\end{bmatrix}\|^2 \\
&= \|\nabla f(x) - \nabla f(x')\|^2 + \|\nabla g(y) - \nabla g(y')\|^2 \\
&\quad \|A^\top(y-y')\|^2 + \|A(x-x')\|^2 \\
&\overset{(i)}{\le} \left(L_f^2 + \|A\|^2\right)\|x-x'\|^2 + \left(L_g^2 + \|A\|^2\right)\|y-y'\|^2 \\
&\le \left(\max\{L_f, L_g\}^2 + \|A\|^2\right)\|\begin{bmatrix}x-x'\\y-y'\end{bmatrix}\|^2 \\
&\le \frac{1}{\lambda_{\min}(\Phi_\eta)}\left(\max\{L_f, L_g\}^2 + \|A\|^2\right)\|\begin{bmatrix}x-x'\\y-y'\end{bmatrix}\|^2_{\Phi_\eta},
\end{aligned}$$

where we used the fact that f and g are smooth in (i). □

Using the previous lemmas, the following contractivity result is straightforward.

Theorem 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L_f smooth. Let $g : \mathbb{R}^m \to \mathbb{R}$ be L_g smooth and μ_g strongly convex. Let $\mu_A = \lambda_{\min}\left(A^\top A\right) > 0$. Let $\eta < \min\left\{\frac{1}{\|A\|}, C_M\right\}$ with C_M as in (34), $\Phi_\eta \succ 0$ as in (32), μ_η and L_η as in Lemma 12 and Lemma 13, respectively. Then, for any $0 < \alpha < \frac{2\mu_\eta}{L_\eta^2}$, Algorithm 4 is contractive in $\|\cdot\|_{\Phi_\eta}$ with rate*

$$\rho = \sqrt{1 - 2\alpha\mu_\eta + \alpha^2 L_\eta^2} < 1. \tag{36}$$

Thus, for all $k \in \mathbb{N}$,

$$\|\begin{bmatrix}x^k - x^\star\\ y^k - y^\star\end{bmatrix}\|_{\Phi_\eta} \le \rho^k \|\begin{bmatrix}x^0 - x^\star\\ y^0 - y^\star\end{bmatrix}\|_{\Phi_\eta},$$

where $(x^\star, y^\star)$ is the unique solution to problem (1).

Proof. Clearly, any fixed point of Algorithm 4 is a solution to (1). The theorem then readily follows in view of Lemma 12 and Lemma 13, see for instance (Bauschke and Combettes, 2017, Prop. 26.16(ii)). □

Algorithm 5 Preconditioned primal-dual gradient method

Require: $x^0 \in \mathbb{R}^n, y^0 \in \mathbb{R}^m$, step sizes $\alpha > 0, \eta > 0$
1: **for** $k = 0, 1, \cdots$ **do**
2: $\quad a^k = \nabla f(x^k) + A^\top y^k$
3: $\quad b^k = \nabla g(y^k) - Ax^k$
4: $\quad x^{k+1} = x^k - \alpha\left(a^k - \eta A^\top b^k\right)$
5: $\quad y^{k+1} = y^k - \alpha\left(b^k - \eta A a^k\right)$
6: **end for**

D.2 Preconditioned Gradient Descent-Ascent

Lemma 12 shows that saddle point operator F is strongly monotone in $\|\cdot\|_{\Phi_\eta}$, when C2 holds and f is smooth. This is an interesting result, and it paves the way for several extensions. For instance, from the proof of Lemma 12, it is immediate to see that, for all $\omega, \omega' \in \mathbb{R}^{m+n}$

$$\begin{aligned}\langle \Phi_\eta F(\omega) - \Phi_\eta F(\omega'), \omega - \omega' \rangle &= \langle F(\omega) - F(\omega'), \omega - \omega' \rangle_{\Phi_\eta} \\ &\geq \lambda_{\min}(M_\eta)\|\omega - \omega'\|^2,\end{aligned}$$

namely, $\Phi_\eta F$ is strongly monotone in the (unweighted) inner product $\langle \cdot, \cdot \rangle$. This suggests to use $\Phi_\eta F$ in the forward step, resulting in Algorithm 5. This algorithm is novel to the best of our knowledge; as for (3), it requires two sequential updates for each iteration, and it is related to the (accelerated) algorithm in Kovalev et al. (2022). Its peculiarity is to be contractive in the unweighted norm $\|\cdot\|$.

Theorem 6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L_f smooth. Let $g : \mathbb{R}^m \to \mathbb{R}$ be L_g smooth and μ_g strongly convex. Let $\mu_A = \lambda_{\min}\left(A^\top A\right) > 0$. Let $\eta < \min\left\{\frac{1}{\|A\|}, C_M\right\}$ with C_M as in (34), $\Phi_\eta \succ 0$ as in (32), μ_η and L_η as in Lemma 12 and Lemma 13, respectively, M_η as in (33). Let $\alpha < \frac{2\lambda_{\min}(M_\eta)}{(L^2+\|A\|^2)\lambda^2_{\max}(\Phi_\eta)}$. Then Algorithm 5 is contractive in $\|\cdot\|$ with rate*

$$\rho = \sqrt{1 - 2\alpha\lambda_{\min}\left(M_\eta\right) + \alpha^2\lambda^2_{\max}(\Phi_\eta)\left(\max\{L_f, L_g\}^2 + \|A\|^2\right)} < 1.$$

Thus, for all $k \in \mathbb{N}$,

$$\| \begin{bmatrix} x^k - x^\star \\ y^k - y^\star \end{bmatrix} \| \leq \rho^k \| \begin{bmatrix} x^0 - x^\star \\ y^0 - y^\star \end{bmatrix} \|,$$

where $(x^\star, y^\star)$ is the unique solution to (1).

Proof. Algorithm 5 can be written as $\omega^{k+1} = (\mathrm{Id} - \alpha\Phi_\eta F)(\omega^k)$. Since $\Phi_\eta \succ 0$, the fixed points of Algorithm 5 coincide with the zeros of F, equivalently withe the solutions to (1). The result follows because the operator $\Phi_\eta F$ is strongly monotone and Lipschitz continuous (in $\|\cdot\|$), see for instance (Bauschke and Combettes, 2017, Prop. 26.16(ii)). ∎