On the ergodic convergence rates of a first-order primal-dual algorithm.

Antonin Chambolle, Thomas Pock

▶ **To cite this version:**

Antonin Chambolle, Thomas Pock. On the ergodic convergence rates of a first-order primal-dual algorithm.. Mathematical Programming, Series A, 2016, 159 (1-2), pp.253-287. hal-01151629v2

HAL Id: hal-01151629
https://hal.science/hal-01151629v2

Submitted on 8 Oct 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the ergodic convergence rates of a first-order primal-dual algorithm

Antonin Chambolle* and Thomas Pock†

September 28, 2015

Abstract

We revisit the proofs of convergence for a first order primal-dual algorithm for convex optimization which we have studied a few years ago. In particular, we prove rates of convergence for a more general version, with simpler proofs and more complete results. The new results can deal with explicit terms and nonlinear proximity operators in spaces with quite general norms.

MSC Classification: 49M29 65K10 65Y20 90C25

Keywords: Saddle-point problems, first order algorithms, primal-dual algorithms, convergence rates, ergodic convergence.

1 Introduction

In this work we revisit a first-order primal-dual algorithm which was introduced in [26, 15] and its accelerated variants which were studied in [5]. We derive new estimates for the rate of convergence. In particular, exploiting a proximal-point interpretation due to [16], we are able to give a very elementary proof of an ergodic $O(1/N)$ rate of convergence (where N is the number of iterations), which also generalizes to non-linear norms [18], to overrelaxed [16, 9] and inertial [19] variants. In the second part, we give new, more precise estimates of the convergence rate for the accelerated variants of the algorithm. We conclude the paper by showing the practical performance of the algorithm on a number of randomly generated standard optimization problems.

The new proofs we propose easily incorporate additional smooth terms such as considered in [9, 31] (where convergence is already been proved, without rates), and [4] (where the proofs of [5] are extended to the framework of [31] which considers general monotone operators—in this setting one must also mention the recent work [10] for a Douglas-Rachford approach to the same problem, with a slightly different algorithm also presenting very good convergence

*CMAP, Ecole Polytechnique, CNRS, 91128 Palaiseau, France.
e-mail: antonin.chambolle@cmap.polytechnique.fr

†Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austriaand Digital Safety & Security Department, AIT Austrian Institute of Technology GmbH, 1220 Vienna, Austria. e-mail: pock@icg.tugraz.at

properties). Also, a very recent work of Drori, Sabach and Teboulle, establishes similar results on a closely related primal-dual ("PAPC") algorithm [11], which also handles explicit terms, but cannot jointly handle (without further splitting) nonsmooth functions in both the primal and dual variables.

We must observe that in addition, our proofs carry on to the nonlinear (or Banach space) setting. They can indeed take into account without effort non-linear proximity operators, based on Bregman distance functions (except in the accelerated variables of the accelerated schemes), in the spirit of the "Mirror-descent" methods introduced by Nemirovski and Yudin [21]. These were extensively studied by many authors, see in particular [29, 6, 2], and [20] in a primal-dual framework. See also [8, 25] for recent advances on such primal-dual algorithms, including stochastic versions. On the other hand, in the standard Euclidean setting, the algorithm we study can be shown to be a particular linearized variant of the ADMM algorithm for which a convergence theory, with more precise results, is found in [28]. We should add that the relationship between the type of algorithms which we study here and the ADMM was already stressed in [5] and that, in particular, one can derive from the analysis in [5] and in this paper convergence rates for the ADMM which are different from the ones currently found in the literature, see for instance [17].

We are addressing the following problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \mathcal{L}(x,y) = \langle Kx, y\rangle + f(x) + g(x) - h^*(y)\,, \tag{1}$$

which is the convex-concave saddle-point form of the "primal" minimization problem

$$\min_{x\in\mathcal{X}} f(x) + g(x) + h(Kx)\,. \tag{2}$$

Here, $\mathcal{X}$ and $\mathcal{Y}$ are, in the most general setting, real reflexive Banach spaces endowed with corresponding norms $\|\cdot\|_x$ and $\|\cdot\|_y$. Note however that in this setting it is quite restrictive to assume that K is bounded, so that the reader could assume that they are finite-dimensional. The only point where it matters is the fact that the estimates we compute never involve the dimension of the current spaces, except possibly through quantities such as $\|K\|$. For notational simplicity, we will drop the subscript for the norms whenever there is no ambiguity. The dual spaces (spaces of all continuous linear functionals) are denoted by $\mathcal{X}^*$, and $\mathcal{Y}^*$. For $x^*\in\mathcal{X}^*$ and $x\in\mathcal{X}$, the bilinear form $\langle x^*, x\rangle$ gives the value of the function x^* at x. Similar, for $y^*\in\mathcal{Y}^*$ and $y\in\mathcal{Y}$, $\langle y^*, y\rangle$ gives the value of the function y^* at y. The norms of the dual spaces are defined as

$$\|x^*\|_* = \sup_{\|x\|\le 1} \langle x^*, x\rangle\,, \quad \|y^*\|_* = \sup_{\|y\|\le 1} \langle y^*, y\rangle\,.$$

By definition, we also have that

$$\langle x^*, x\rangle \le \|x\|\cdot\|x^*\|_*, \quad \langle y^*, y\rangle \le \|y\|\cdot\|y^*\|_*.$$

We further assume that the following assumptions are fulfilled:

(i) $K:\mathcal{X}\to\mathcal{Y}^*$ is a bounded linear operator, with corresponding adjoint operator $K^*:\mathcal{Y}\to\mathcal{X}^*$ defined by

$$\langle Kx, y\rangle = \langle K^*y, x\rangle \quad \forall (x,y)\in\mathcal{X}\times\mathcal{Y}.$$

Throughout the whole paper we will keep the notation "L" for the norm of this operator, defined by

$$L := \|K\| \;=\; \sup_{\|x\|\leq 1,\, \|y\|\leq 1} \langle Kx, y\rangle \;=\; \sup_{\|x\|\leq 1} \|Kx\|_* = \|K^*\| \;=\; \sup_{\|y\|\leq 1} \|K^* y\|_*.$$

Hence, we also have that

$$\langle Kx, y\rangle \leq \|Kx\|_* \|y\| \leq L\,\|x\|\,\|y\|, \qquad \langle K^* y, x\rangle \leq \|K^* y\|_* \|x\| \leq L\,\|x\|\,\|y\|.$$

For example, let $\|\cdot\|_x = \|\cdot\|_p$ and $\|\cdot\|_y = \|\cdot\|_q$, with $p, q \geq 1$, i.e. the usual ℓ_p norms, then

$$\|K\| = \sup_{\|x\|_p\leq 1} \|Kx\|_{q'} = \sup_{\|y\|_q\leq 1} \|K^* y\|_{p'} = \sup_{\substack{\|x\|_p\leq 1\\ \|y\|_q\leq 1}} \langle Kx, y\rangle\,,$$

with p', q' such that $1/p + 1/p' = 1$, and $1/q + 1/q' = 1$.

(ii) f is a proper, lower semicontinuous (l.s.c.), convex function, with ∇f Lipschitz continuous on $\mathcal{X}$, i.e.

$$\|\nabla f(x) - \nabla f(x')\|_* \leq L_f \|x - x'\|\,, \quad \forall x, x' \in \mathcal{X};$$

(iii) g, h are proper, l.s.c., convex functions with simple structure, in the sense that their proximal maps

$$\min_x g(x) + \frac{1}{\tau} D_x(x, \bar{x})\,, \quad \min_y h^*(y) + \frac{1}{\sigma} D_y(y, \bar{y})\,,$$

can be computed for any $\tau, \sigma > 0$.

Here D_x and D_y are Bregman proximity/distance functions based on 1-strongly convex (w.r.t. the respective norms) functions ψ_x and ψ_y, defined by

$$D_x(x, \bar{x}) = \psi_x(x) - \psi_x(\bar{x}) - \langle \nabla\psi_x(\bar{x}), x - \bar{x}\rangle\,,$$
$$D_y(y, \bar{y}) = \psi_y(y) - \psi_y(\bar{y}) - \langle \nabla\psi_y(\bar{y}), y - \bar{y}\rangle\,.$$

Following [13], we assume that ψ_x, ψ_y are continuously differentiable on open sets S_x, S_y, continuous on $\overline{S}_x$, $\overline{S}_y$, and that given any converging sequences (x^n) and (y^n),

$$x^n \to x \Rightarrow \lim_{n\to\infty} D_x(x, x^n) = 0, \quad y^n \to y \Rightarrow \lim_{n\to\infty} D_y(y, y^n) = 0. \tag{3}$$

We may of course assume that $\overline{S}_x$ and $\overline{S}_y$ are the respective domains of ψ_x, ψ_y. We need, in addition to [13], to assume the strong convexity of our functions to ensure the convergence of the algorithms studied in this paper. This restricts the possible class of Bregman functions, notice however that classical examples such as the entropy $\psi_x(x) = \sum_{i=1}^d x_i \log x_i$ is well-known to be 1-strongly convex with respect to the 1-norm [2, 29] when restricted to the unit simplex, it is also strongly convex with respect to the 2-norm on bounded sets of $(\mathbb{R}_+)^d$. Eventually, we must assume here that $\operatorname{dom} g \subseteq \operatorname{dom} \psi_x = \overline{S}_x$ and $\operatorname{dom} h^* \subseteq \operatorname{dom} \psi_y = \overline{S}_y$.

Clearly, the Lipschitz continuity of f implies that

$$f(x') \leq f(x) + \langle \nabla f(x), x' - x\rangle + \frac{L_f}{2}\|x' - x\|^2\,, \quad \forall x, x' \in \mathcal{X}\,. \tag{4}$$

Furthermore, the 1-strongy convexity of ψ_x and ψ_y easily implies that for any $x, \bar{x}$ and $y, \bar{y}$, it holds

$$D_x(x, \bar{x}) \geq \tfrac{1}{2}\|x - \bar{x}\|^2 , \quad D_y(y, \bar{y}) \geq \tfrac{1}{2}\|y - \bar{y}\|^2 .$$

The most common choice for ψ_x and ψ_y is the usual squared Euclidean norm $\frac{1}{2}\|\cdot\|_2^2$ (or Hilbertian in infinite dimension), which yields

$$D(x, \bar{x}) = \frac{1}{2}\|x - \bar{x}\|_2^2 .$$

We will refer to this classical case as the "Euclidean case" . In this case, it is standard that given a convex, l.s.c. function ϕ, if $\hat{u}$ is the minimizer of

$$\phi(u) + \frac{1}{2}\|u - \bar{u}\|_2^2$$

(which we call the "Euclidean proximity map" of ϕ at $\bar{u}$), then by strong convexity one has for all u

$$\phi(u) + \frac{1}{2}\|u - \bar{u}\|_2^2 \geq \phi(\hat{u}) + \frac{1}{2}\|\hat{u} - \bar{u}\|_2^2 + \frac{1}{2}\|u - \hat{u}\|_2^2 .$$

It turns out that this property is true also for non-Euclidean proximity operators, that is

$$\hat{u} = \arg\min_u \phi(u) + D(u, \bar{u}) \quad \Longrightarrow \quad \forall u,\ \phi(u) + D(u, \bar{u}) \geq \phi(\hat{u}) + D(\hat{u}, \bar{u}) + D(u, \hat{u}). \tag{5}$$

This is easily deduced from the optimality conditions for $\hat{u}$, see [6, 30].

Before closing this section, we point out that most of our results still hold, if the function h is a convex l.s.c. function of the form [31, 4, 19]

$$h(y) = \min_{y_1 + y_2 = y} h_1(y_1) + h_2(y_2), \tag{6}$$

so that

$$h^*(y) = h_1^*(y) + h_2^*(y),$$

h_1^* having simple structure while ∇h_2^* can be evaluated and is Lipschitz continuous with parameter $L_{h_2^*}$. For the ease of presentation we will not consider this situation but we will mention when our results can be extended to this case.

2 The general iteration

Iteration: $(\hat{x}, \hat{y}) = \mathcal{PD}_{\tau,\sigma}(\bar{x}, \bar{y}, \tilde{x}, \tilde{y})$

$$\begin{cases} \hat{x} = \arg\min_x f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + g(x) + \langle Kx, \tilde{y}\rangle + \frac{1}{\tau} D_x(x, \bar{x}) \\ \hat{y} = \arg\min_y h^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{\sigma} D_y(y, \bar{y}) . \end{cases} \tag{7}$$

The main iterate of the class of primal-dual algorithms we consider in this paper is defined in (7). It takes the points $(\bar{x}, \bar{y})$ as well as the intermediate points $(\tilde{x}, \tilde{y})$ as input and outputs the new points $(\hat{x}, \hat{y})$. It satisfies the following descent rule:

Lemma 1. *If* (7) *holds, then for any* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$ *one has*

$$\begin{aligned}\mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \hat{y}) \leq \frac{1}{\tau} D_x(x, \bar{x}) - \frac{1}{\tau} D_x(x, \hat{x}) - \frac{1}{\tau} D_x(\hat{x}, \bar{x}) + \frac{L_f}{2}\|\hat{x} - \bar{x}\|^2 \\ + \frac{1}{\sigma} D_y(y, \bar{y}) - \frac{1}{\sigma} D_y(y, \hat{y}) - \frac{1}{\sigma} D_y(\hat{y}, \bar{y}) + \\ + \langle K(x - \hat{x}), \tilde{y} - \hat{y}\rangle - \langle K(\tilde{x} - \hat{x}), y - \hat{y}\rangle \,. \quad (8)\end{aligned}$$

Proof. From the first line in the above iteration (7) and property (5), it follows:

$$\begin{aligned}\langle \nabla f(\bar{x}), x\rangle + g(x) + \langle Kx, \tilde{y}\rangle + \tfrac{1}{\tau} D_x(x, \bar{x}) \geq \\ \langle \nabla f(\bar{x}), \hat{x}\rangle + g(\hat{x}) + \langle K\hat{x}, \tilde{y}\rangle + \tfrac{1}{\tau} D_x(\hat{x}, \bar{x}) + \tfrac{1}{\tau} D_x(x, \hat{x}) \,.\end{aligned}$$

Moreover, from the convexity of f and (4) it follows

$$f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle \geq f(\hat{x}) + \langle \nabla f(\bar{x}), x - \hat{x}\rangle - \frac{L_f}{2}\|\hat{x} - \bar{x}\|^2 \,.$$

Combining this with the previous inequality, we arrive at

$$\begin{aligned}f(x) + g(x) + \tfrac{1}{\tau} D_x(x, \bar{x}) + \frac{L_f}{2}\|\hat{x} - \bar{x}\|^2 \geq \\ f(\hat{x}) + g(\hat{x}) + \langle K(\hat{x} - x), \tilde{y}\rangle + \tfrac{1}{\tau} D_x(\hat{x}, \bar{x}) + \tfrac{1}{\tau} D_x(x, \hat{x}) \,. \quad (9)\end{aligned}$$

In the same way:

$$h^*(y) + \tfrac{1}{\sigma} D_y(y, \bar{y}) \geq h^*(\hat{y}) - \langle K\tilde{x}, \hat{y} - y\rangle + \tfrac{1}{\sigma} D_y(\hat{y}, \bar{y}) + \tfrac{1}{\sigma} D_y(y, \hat{y}) \,. \quad (10)$$

Summing (9), (10) and rearranging the terms appropriately, we obtain (8). □

3 Non-linear primal-dual algorithm

In this section we address the convergence rate of the non-linear primal-dual algorithm shown in Algorithm 1: The elegant interpretation in [16] shows that by writing the algorithm in this form

Algorithm 1: $O(1/N)$ Non-linear primal-dual algorithm

- Input: Operator norm $L := \|K\|$, Lipschitz constant L_f of ∇f, and Bregman distance functions D_x and D_y.
- Initialization: Choose $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$, $\tau, \sigma > 0$
- Iterations: For each $n \geq 0$ let

$$(x^{n+1}, y^{n+1}) = \mathcal{PD}_{\tau,\sigma}(x^n, y^n, 2x^{n+1} - x^n, y^n) \quad (11)$$

(which "shifts" the updates with respect to [5]), in the Euclidean case, that is $\|\cdot\|_x = \|\cdot\|_y = \|\cdot\|_2$,

and $D_x(x,x') = \frac{1}{2}\|x-x'\|_2^2$, $D_y(y,y') = \frac{1}{2}\|y-y'\|_2^2$, then it is an instance of the *proximal point algorithm* [27], up to the explicit term $\nabla f(x^n)$, since

$$\begin{pmatrix} K^* + \partial g \\ -K + \partial h^* \end{pmatrix} (z^{n+1}) + M_{\tau,\sigma}(z^{n+1} - z^n) \ni \begin{pmatrix} -\nabla f(x^n) \\ 0 \end{pmatrix},$$

where the variable $z \in \mathcal{X} \times \mathcal{Y}$ represents the pair (x,y), and the matrix $M_{\tau,\sigma}$ is given by

$$M_{\tau,\sigma} = \begin{pmatrix} \frac{1}{\tau} I & -K^* \\ -K & \frac{1}{\sigma} I \end{pmatrix}, \tag{12}$$

which is positive-definite as soon as $\tau\sigma L^2 < 1$. A proof of convergence is easily deduced. Moreover, since in our particular setting we never really use the machinery of monotone operators, and rely only on the fact that we are studying a specific saddle-point problem, the results are a bit improved: in particular we deal easily with the explicit term f and non-linear proximity operators.

Theorem 1. *Let (x^n, y^n), $n = 0, \dots, N-1$ be a sequence generated by the non-linear primal-dual algorithm* (11)*. Let the step size parameters $\tau, \sigma > 0$ be chosen such that for all $x, x' \in \operatorname{dom} g$ and $y, y' \in \operatorname{dom} h^*$ it holds that*

$$\left(\frac{1}{\tau} - L_f\right) D_x(x,x') + \frac{1}{\sigma} D_y(y,y') - \langle K(x-x'), y-y' \rangle \geq 0\,. \tag{13}$$

Then, for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$ it holds that

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \leq \frac{1}{N}\left(\frac{1}{\tau} D_x(x,x^0) + \frac{1}{\sigma} D_y(y,y^0) - \left\langle K(x-x^0), y-y^0 \right\rangle\right), \tag{14}$$

where $X^N = \frac{1}{N}\sum_{n=1}^N x^n$, and $Y^N = \frac{1}{N}\sum_{n=1}^N y^n$.

Proof. According to the iterative scheme (11), the estimate (8) becomes

$$\begin{aligned} \mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \leq{}& \left[\frac{1}{\tau} D_x(x,x^n) + \frac{1}{\sigma} D_y(y,y^n) - \langle K(x-x^n), y-y^n \rangle\right] \\ &- \left[\frac{1}{\tau} D_x(x,x^{n+1}) + \frac{1}{\sigma} D_y(y,y^{n+1}) - \left\langle K(x-x^{n+1}), y-y^{n+1} \right\rangle\right] \\ &- \left[\frac{1}{\tau} D_x(x^{n+1},x^n) + \frac{1}{\sigma} D_y(y^{n+1},y^n) - \left\langle K(x^{n+1}-x^n), y^{n+1}-y^n \right\rangle - \frac{L_f}{2}\|x^{n+1}-x^n\|^2\right]. \end{aligned} \tag{15}$$

Thanks to (13), the terms in the brackets are non-negative. Now we sum the last estimate from $n = 0, \dots, N-1$ and find

$$\sum_{n=1}^N \mathcal{L}(x^n, y) - \mathcal{L}(x, y^n) \leq \frac{1}{\tau} D_x(x,x^0) + \frac{1}{\sigma} D_y(y,y^0) - \left\langle K(x-x^0), y-y^0 \right\rangle,$$

where we have removed negative terms on the right hand side. Equation (14) follows from the convexity of $(\xi,\eta) \mapsto \mathcal{L}(\xi, y) - \mathcal{L}(x, \eta)$. □

Remark 1. Observe that since $D_x(\cdot, x')$ and $D_y(\cdot, y')$ are 1-convex, (13) is ensured as soon as

$$\left(\frac{1}{\tau} - L_f\right)\frac{1}{\sigma} \geq L^2. \tag{16}$$

Remark 2. The rate (14) can also be written solely in terms of the distance functions D_x and D_y. In fact, for any $\alpha > 0$,

$$|\left\langle K(x-x^0), y-y^0\right\rangle| \leq L\|x-x^0\|\|y-y^0\| \leq \\ \frac{\alpha L}{2}\|x-x^0\|^2 + \frac{L}{2\alpha}\|y-y^0\|^2 \leq \alpha L D_x(x,x^0) + \frac{L}{\alpha}D_y(y,y^0).$$

In case $L_f = 0$, $\tau\sigma L^2 = 1$ and choosing $\alpha = 1/(\tau L)$, the rate (14) becomes

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \leq \frac{2}{N}\left(\frac{1}{\tau}D_x(x,x^0) + \frac{1}{\sigma}D_y(y,y^0)\right). \tag{17}$$

In the Euclidean setting, that is $\|\cdot\|_x = \|\cdot\|_y = \|\cdot\|_2 = \langle\cdot,\cdot\rangle^{\frac{1}{2}}$, and $D(x,x') = \frac{1}{2}\|x-x'\|_2^2$, $D(y,y') = \frac{1}{2}\|y-y'\|_2^2$, the estimate (15) reduces to

$$\mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \leq \frac{1}{2}\|z-z^n\|^2_{M_{\tau,\sigma}} - \frac{1}{2}\|z-z^{n+1}\|^2_{M_{\tau,\sigma}} \\ - \frac{1}{2}\|z^{n+1}-z^n\|^2_{M_{\tau,\sigma}} + \frac{L_f}{2}\|x^{n+1}-x^n\|_2^2,$$

with $M_{\tau,\sigma}$ defined in (12). This can also be rewritten as

$$\mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \leq \left\langle z^{n+1}-z^n, z-z^{n+1}\right\rangle_{M_{\tau,\sigma}} + \frac{L_f}{2}\|x^{n+1}-x^n\|_2^2 \tag{18}$$

while the final estimate (14) becomes

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \leq \frac{1}{2N}\|z-z^0\|^2_{M_{\tau,\sigma}}. \tag{19}$$

Observe that this rate is different from the rate obtained in [5], which does only depend on the diagonal part of $M_{\tau,\sigma}$ (each rate can be bounded by twice the other).

Remark 3. If we assume in addition that the inequality $\tau\sigma L^2 < 1$ is strict (which follows from (16) if $L_f > 0$, and has to be assumed else), then we can deduce as in [5] convergence results for the algorithm, whenever a saddle-point $z^* = (x^*, y^*)$ exists. The first thing to observe is that this inequality yields that

$$\frac{1}{\tau}D_x(x,x') + \frac{1}{\sigma}D_y(y,y') - \langle K(x-x'), y-y'\rangle \geq \alpha(\|x-x'\|^2 + \|y-y'\|^2) \tag{20}$$

for some $\alpha > 0$. As a consequence, it follows from (15) that the sequence $z^n = (x^n, y^n)$ is globally bounded (indeed, $\mathcal{L}(X^N, y^*) - \mathcal{L}(x^*, Y^N) \geq 0$). Obviously, this also yields a bound for $Z^N = (X^N, Y^N)$. We may thus assume that a subsequence $(Z^{N_k})_k$ weakly converges in $\mathcal{X}\times\mathcal{Y}$ to some $Z = (X, Y)$, and from (14) and the lower-semicontinuity of f, g, h^* it follows that the limit Z is a saddle-point.

In finite dimension, we can also show the convergence of the whole sequences z^n and Z^N to the same saddle-point. The proof follows the proof in [26, 5], in the linear case. Let us assume that z is a limit point for a subsequence $(z^{n_k})_k$, then since (15) guaranties the summability of $\|z^{n+1} - z^n\|^2$, we have that also $z^{n_k \pm 1} \to z$. It follows that z is a fixed point of the algorithm and thus a saddle-point (which we now denote $z^* = (x^*, y^*)$).

Let $m \geq 0$ be the limit of the nonincreasing sequence

$$\frac{1}{\tau} D_x(x^*, x^n) + \frac{1}{\sigma} D_y(y^*, y^n) - \langle K(x^* - x^n), y^* - y^n \rangle ,$$

we wish to show that $m = 0$. Since $z^{n_k} \to z^*$ we deduce

$$\lim_{k \to \infty} \frac{1}{\tau} D_x(x^*, x^{n_k}) + \frac{1}{\sigma} D_y(y^*, y^{n_k}) = m.$$

Using assumption (3), we deduce $m = 0$. The convergence of the global sequence follows from (20). In Hilbert spaces of infinite dimension, the same proof shows weak convergence of the sequence for Euclidean proximity operators, invoking Opial's theorem [24].

Remark 4. In the Euclidean setting and when $g = 0$, a better algorithm (in fact, optimal, see [21, 23]) is proposed in [7], which yields a rate of order $O(L_f/N^2 + L/N)$.

Remark 5. In case h has the composite form (6), then the theorem still holds with the condition (16) replaced with

$$\left(\frac{1}{\tau} - L_f \right) \left(\frac{1}{\sigma} - L_{h_2^*} \right) \geq L^2. \tag{21}$$

4 Overrelaxed and inertial variants

In this section, we consider overrelaxed and inertial versions of the primal-dual algorithm. We will only consider the Euclidean setting, that is $\|\cdot\|_x = \|\cdot\|_y = \|\cdot\|_2 = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$, and $D(x, x') = \frac{1}{2}\|x - x'\|_2^2$, $D(y, y') = \frac{1}{2}\|y - y'\|_2^2$, since our proofs heavily rely on the fact that $\|\cdot\|_2^2 = \langle \cdot, \cdot \rangle$.

4.1 Relaxed primal-dual algorithm

Algorithm 2: $O(1/N)$ Overrelaxed primal-dual algorithm

- Input: Operator norm $L = \|K\|_{2,2}$, Lipschitz constant L_f of ∇f, Bregman distance functions $D(x, x') = \frac{1}{2}\|x - x'\|_2^2$, $D(y, y') = \frac{1}{2}\|y - y'\|_2^2$.
- Initialization: Choose $z^0 = (x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$, $\tau, \sigma > 0$ and $\rho_n \in (0, 2)$
- Iterations: For each $n \geq 0$ let

$$\begin{cases} (\xi^{n+1}, \eta^{n+1}) = \mathcal{PD}_{\tau,\sigma}(x^n, y^n, 2\xi^{n+1} - x^n, y^n) \\ z^{n+1} = (1 - \rho_n) z^n + \rho_n \zeta^{n+1} \end{cases} \tag{22}$$

where $z^n = (x^n, y^n)$ and $\zeta^n = (\xi^n, \eta^n)$.

First we consider the overrelaxed primal-dual Algorithm 2, whose convergence has been considered already in [14, 16]. It is known that an overrelaxation parameter close to 2 can speed up the convergence but a theoretical justification was still missing.

Theorem 2. *Assume* $\|\cdot\|_x = \|\cdot\|_y = \|\cdot\|_2 = \langle\cdot,\cdot\rangle^{\frac{1}{2}}$, $D_x(x,x') = \frac{1}{2}\|x-x'\|_2^2$, $D_y(y,y') = \frac{1}{2}\|y-y'\|_2^2$. *Let* (ξ^n,η^n), $n=0,\dots,N-1$ *be a sequence generated by the overrelaxed Euclidean primal-dual algorithm* (22). *Let the step size parameters* $\tau,\sigma>0$ *and the overrelaxation parameter* ρ_n *be a non-decreasing sequence in* $(0,\rho)$ *with* $\rho<2$ *such that for all* $x,x'\in\operatorname{dom} g$ *and* $y,y'\in\operatorname{dom} h^*$ *it holds that*

$$\left(\frac{1}{\tau}-\frac{L_f}{2-\rho}\right)\frac{1}{\sigma} > \|K\|_2^2. \tag{23}$$

Then, for any $z=(x,y)\in\mathcal{X}\times\mathcal{Y}$ *it holds that*

$$\mathcal{L}(X^N,y)-\mathcal{L}(x,Y^N) \le \frac{1}{2\rho_0 N}\|z-z^0\|_{M_{\tau,\sigma}}^2, \tag{24}$$

where $X^N=\frac{1}{N}\sum_{n=1}^N\xi^n$, *and* $Y^N=\frac{1}{N}\sum_{n=1}^N\eta^n$.

Proof. We start with the basic inequality (8). According to (22), using $\bar z=z^n$ and $\tilde z=(2\xi^{n+1}-x^n,y^n)$ and $\hat z=\zeta^{n+1}$, we obtain

$$\mathcal{L}(\xi^{n+1},y)-\mathcal{L}(x,\eta^{n+1}) \le \left\langle \zeta^{n+1}-z^n, z-\zeta^{n+1}\right\rangle_{M_{\tau,\sigma}} + \frac{L_f}{2}\|\xi^{n+1}-x^n\|_2^2,$$

where $M_{\tau,\sigma}$ is defined in (12) and we have used the fact that $2\langle a,b\rangle_M = \|a\|_M^2+\|b\|_M^2-\|a-b\|_M^2$. Now, observe that from the second line in (22), the auxiliary point ζ^{n+1} can be written as

$$\zeta^{n+1} = z^n + \frac{1}{\rho_n}(z^{n+1}-z^n).$$

Substituting back into the previous inequality, we have

$$\begin{aligned}
&\mathcal{L}(\xi^{n+1},y)-\mathcal{L}(x,\eta^{n+1})\\
&\quad\le \left\langle z^n+\frac{1}{\rho_n}(z^{n+1}-z^n)-z^n, z-z^n-\frac{1}{\rho_n}(z^{n+1}-z^n)\right\rangle_{M_{\tau,\sigma}}\\
&\qquad\qquad + \frac{L_f}{2}\|x^n+\frac{1}{\rho_n}(x^{n+1}-x^n)-x^n\|_2^2\\
&\quad= \frac{1}{\rho_n}\left\langle z^{n+1}-z^n, z-z^n\right\rangle_{M_{\tau,\sigma}} - \frac{1}{\rho_n^2}\|z^{n+1}-z^n\|_{M_{\tau,\sigma}}^2 + \frac{L_f}{2\rho_n^2}\|x^{n+1}-x^n\|_2^2\\
&\quad= \frac{1}{2\rho_n}\left(\|z-z^n\|_{M_{\tau,\sigma}}^2 - \|z-z^{n+1}\|_{M_{\tau,\sigma}}^2\right)\\
&\qquad\qquad - \frac{2-\rho_n}{2\rho_n^2}\|z^{n+1}-z^n\|_{M_{\tau,\sigma}}^2 + \frac{L_f}{2\rho_n^2}\|x^{n+1}-x^n\|_2^2\\
&\quad\le \frac{1}{2\rho_n}\left(\|z-z^n\|_{M_{\tau,\sigma}}^2 - \|z-z^{n+1}\|_{M_{\tau,\sigma}}^2\right) - \frac{2-\rho_n}{2\rho_n^2}\|z^{n+1}-z^n\|_{M_{\tau,\sigma,\rho_n}}^2,
\end{aligned}$$

where we have defined the metric

$$M_{\tau,\sigma,\rho_n} = \begin{pmatrix} (\frac{1}{\tau}-\frac{L_f}{2-\rho_n})I & -K^*\\ -K & \frac{1}{\sigma}I\end{pmatrix},$$

which is positive definite for all n as soon as (23) is fulfilled. In addition, since ρ_n is a non-decreasing sequence in $(0, \rho)$ with $\rho < 2$, summing the above inequality for $n = 0, \dots, N-1$ and omitting all nonpositive terms on the right hand side, it follows

$$\sum_{n=1}^{N} \mathcal{L}(\xi^n, y) - \mathcal{L}(x, \eta^n) \leq \frac{1}{2\rho_0} \|z - z^0\|^2_{M_{\tau,\sigma}}.$$

The final estimate (24) follows from defining appropriate averages and the convexity of the gap function. □

Remark 6. The last result indeed shows that the convergence rate is improved by choosing ρ_0 as large as possible, i.e. close to 2. However, observe that in case the smooth explicit term ∇f is not zero, it might be less beneficial to use a overrerlaxation parameter larger than one since it requires a smaller primal step size τ.

4.2 Inertial primal-dual algorithm

Next, we consider an inertial version of the primal-dual algorithm, who has recently been considered in [19] as an extension of the inertial proximal point algorithm of Alvarez and Attouch [1]. It has already been observed in numerical experiments that inertial terms leads to a faster convergence of the algorithm. Here we give a theoretical evidence that indeed the presence of an inertial term leads to a smaller worst-case complexity.

Algorithm 3: $O(1/N)$ Inertial primal-dual algorithm

- Input: Operator norm $L = \|K\|_{2,2}$, Lipschitz constant L_f of ∇f, and Bregman distance functions $D_x(x, x') = \frac{1}{2}\|x - x'\|_2^2$ and $D_y(y, y') = \frac{1}{2}\|y - y'\|_2^2$.
- Initialization: Choose $(x^{-1}, y^{-1}) = (x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$, $\tau, \sigma > 0$ and $\alpha_n \in [0, 1/3)$
- Iterations: For each $n \geq 0$ let[a]

$$\begin{cases} \zeta^n = z^n + \alpha_n(z^n - z^{n-1}) \\ (x^{n+1}, y^{n+1}) = \mathcal{PD}_{\tau,\sigma}(\xi^n, \eta^n, 2x^{n+1} - \xi^n, \eta^n) \end{cases} \tag{25}$$

[a] Here as before, $z = (x, y)$ and similarly, $\zeta = (\xi, \eta)$.

Theorem 3. *Assume* $\|\cdot\|_x = \|\cdot\|_y = \|\cdot\|_2 = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$, $D_x(x, x') = \frac{1}{2}\|x - x'\|_2^2$, $D_y(y, y') = \frac{1}{2}\|y - y'\|_2^2$. *Let* (x^n, y^n), $n = 0, \dots, N-1$ *be a sequence generated by the inertial Euclidean primal-dual algorithm* (25). *Let the step size parameters* $\tau, \sigma > 0$ *and the inertial parameter* α_n *be a non-decreasing sequence in* $[0, \alpha]$ *with* $\alpha < 1/3$ *such that for all* $x, x' \in \operatorname{dom} g$ *and* $y, y' \in \operatorname{dom} h^*$ *it holds that*

$$\left(\frac{1}{\tau} - \frac{(1+\alpha)^2}{1-3\alpha} L_f\right) \frac{1}{\sigma} > \|K\|_2^2. \tag{26}$$

Then, for any $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ *it holds that*

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \leq \frac{1 - \alpha_0}{2N} \|z - z^0\|^2_{M_{\tau,\sigma}}, \tag{27}$$

where $X^N = \frac{1}{N}\sum_{n=1}^N x^n$, *and* $Y^N = \frac{1}{N}\sum_{n=1}^N y^n$.

Proof. We again start with the basic inequality (8). According to (25), using $\bar{z} = \zeta^n$ and $\hat{z} = z^{n+1}$, we have

$$\mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \leq \left\langle z^{n+1} - \zeta^n, z - z^{n+1} \right\rangle_{M_{\tau,\sigma}} + \frac{L_f}{2}\|x^{n+1} - \xi^n\|_2^2 .$$

Plugging in the first line of (25) we arrive at

$$\begin{aligned}
&\mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \\
&\qquad \leq \left\langle z^{n+1} - z^n, z - z^{n+1} \right\rangle_{M_{\tau,\sigma}} - \alpha_n \left\langle z^n - z^{n-1}, z - z^{n+1} \right\rangle_{M_{\tau,\sigma}} \\
&\qquad\qquad + \frac{L_f}{2}\|x^{n+1} - x^n - \alpha_n(x^n - x^{n-1})\|_2^2 \\
&\qquad \leq \left\langle z^{n+1} - z^n, z - z^{n+1} \right\rangle_{M_{\tau,\sigma}} - \alpha_n \left\langle z^n - z^{n-1}, z - z^n + z^n - z^{n+1} \right\rangle_{M_{\tau,\sigma}} \\
&\qquad\qquad + \frac{L_f}{2}\left((1+\alpha_n)\|x^{n+1} - x^n\|_2^2 + (\alpha_n + \alpha_n^2)\|x^n - x^{n-1}\|_2^2\right) \\
&\qquad \leq \frac{1}{2}\left(\|z - z^n\|_{M_{\tau,\sigma}}^2 - \|z - z^{n+1}\|_{M_{\tau,\sigma}}^2 - \|z^{n+1} - z^n\|_{M_{\tau,\sigma}}^2\right) \\
&\qquad\qquad - \frac{\alpha_n}{2}\left(\|z - z^{n-1}\|_{M_{\tau,\sigma}}^2 - \|z - z^n\|_{M_{\tau,\sigma}}^2 - \|z^n - z^{n-1}\|_{M_{\tau,\sigma}}^2\right) \\
&\qquad\qquad - \alpha_n \left\langle z^n - z^{n-1}, z^n - z^{n+1} \right\rangle_{M_{\tau,\sigma}} \\
&\qquad\qquad + \frac{L_f}{2}\left((1+\alpha_n)\|x^{n+1} - x^n\|_2^2 + (\alpha_n + \alpha_n^2)\|x^n - x^{n-1}\|_2^2\right).
\end{aligned}$$

Using the inequality $|\langle a, b\rangle_M| \leq \frac{1}{2}\left(\|a\|_M^2 + \|b\|_M^2\right)$ we obtain the estimate

$$\begin{aligned}
&\mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) \\
&\qquad \leq \frac{1}{2}\left(\|z - z^n\|_{M_{\tau,\sigma}}^2 - \|z - z^{n+1}\|_{M_{\tau,\sigma}}^2\right) + \frac{\alpha_n}{2}\left(\|z - z^n\|_{M_{\tau,\sigma}}^2 - \|z - z^{n-1}\|_{M_{\tau,\sigma}}^2\right) \\
&\qquad\qquad + \frac{\alpha_n - 1}{2}\|z^{n+1} - z^n\|_{M_{\tau,\sigma}}^2 + \alpha_n\|z^n - z^{n-1}\|_{M_{\tau,\sigma}}^2 \\
&\qquad\qquad\qquad + \frac{L_f}{2}\left((1+\alpha_n)\|x^{n+1} - x^n\|_2^2 + (\alpha_n + \alpha_n^2)\|x^n - x^{n-1}\|_2^2\right).
\end{aligned}$$

Now, since $\alpha_n \geq 0$ is non-decreasing and $z^{-1} = z^0$, summing the above inequality for $n = 0, \ldots, N-1$, we find:

$$\begin{aligned}
\sum_{n=1}^N \mathcal{L}(x^n, y) - \mathcal{L}(x, y^n) &\leq \frac{1 - \alpha_0}{2}\|z - z^0\|_{M_{\tau,\sigma}}^2 - \frac{1}{2}\|z - z^N\|_{M_{\tau,\sigma}}^2 \\
&\quad + \frac{\alpha_{N-1}}{2}\|z - z^{N-1}\|_{M_{\tau,\sigma}}^2 + \sum_{n=0}^{N-2}\frac{3\alpha_{n+1} - 1}{2}\|z^{n+1} - z^n\|_{M_{\tau,\sigma,\alpha_{n+1}}}^2 \\
&\quad + \frac{\alpha_{N-1} - 1}{2}\|z^N - z^{N-1}\|_{M_{\tau,\sigma}}^2 + \frac{L_f}{2}(1 + \alpha_{N-1})\|x^N - x^{N-1}\|_2^2,
\end{aligned}$$

where

$$M_{\tau,\sigma,\alpha_n} = \begin{pmatrix} \left(\frac{1}{\tau} - \frac{(1+\alpha_n)^2}{1-3\alpha_n}L_f\right)I & -K^* \\ -K & \frac{1}{\sigma}I \end{pmatrix},$$

which is positive definite for all n as soon as (26) is fulfilled for all $\alpha_n \leq \alpha < 1/3$ since the function $\frac{(1+\alpha_n)^2}{1-3\alpha_n}$ is monotonically increasing in α_n. Our last estimate can be further simplified as

$$\begin{aligned}\sum_{n=1}^{N} \mathcal{L}(x^n, y) - \mathcal{L}(x, y^n) \leq \frac{1-\alpha_0}{2}\|z - z^0\|^2_{M_{\tau,\sigma}} \\ + \frac{\alpha_{N-1}}{2}\|z - z^{N-1}\|^2_{M_{\tau,\sigma}} + \frac{\alpha_{N-1}-1}{2}\|z^N - z^{N-1}\|^2_{M_{\tau,\sigma}} - \frac{1}{2}\|z - z^N\|^2_{M_{\tau,\sigma}} \\ + \frac{L_f}{2}(1+\alpha_{N-1})\|x^N - x^{N-1}\|_2^2\end{aligned}$$

It remains to show that the term in the last two lines of the above estimate is nonpositive. In fact:

$$\begin{aligned}&\frac{\alpha_{N-1}}{2}\|z - z^N + z^N - z^{N-1}\|^2_{M_{\tau,\sigma}} + \frac{\alpha_{N-1}-1}{2}\|z^N - z^{N-1}\|^2_{M_{\tau,\sigma}} \\ &\qquad - \frac{1}{2}\|z - z^N\|^2_{M_{\tau,\sigma}} + \frac{L_f}{2}(1+\alpha_{N-1})\|x^N - x^{N-1}\|_2^2 \\ &\leq \alpha_{N-1}\left(\|z - z^N\|^2_{M_{\tau,\sigma}} + \|z^N - z^{N-1}\|^2_{M_{\tau,\sigma}}\right) + \frac{\alpha_{N-1}-1}{2}\|z^N - z^{N-1}\|^2_{M_{\tau,\sigma}} \\ &\qquad - \frac{1}{2}\|z - z^N\|^2_{M_{\tau,\sigma}} + \frac{L_f}{2}(1+\alpha_{N-1})\|x^N - x^{N-1}\|_2^2 \\ &= (\alpha_{N-1} - \frac{1}{2})\|z - z^N\|^2_{M_{\tau,\sigma}} + \frac{3\alpha_{N-1}-1}{2}\|z^N - z^{N-1}\|^2_{M_{\tau,\sigma}} \\ &\qquad + \frac{L_f}{2}(1+\alpha_{N-1})\|x^N - x^{N-1}\|_2^2 \\ &= (\alpha_{N-1} - \frac{1}{2})\|z - z^N\|^2_{M_{\tau,\sigma}} + \frac{3\alpha_{N-1}-1}{2}\|z^N - z^{N-1}\|^2_{P} \;\leq\; 0,\end{aligned}$$

as $\alpha_n \leq \alpha < 1/3$ and as the matrix

$$P = \begin{pmatrix} (\frac{1}{\tau} - \frac{1+\alpha_{N-1}}{1-3\alpha_{N-1}}L_f)I & -K^* \\ -K & \frac{1}{\sigma}I \end{pmatrix}$$

is clearly positive definite if (26) is fulfilled. It remains to derive the ergodic rate by defining appropriate averages and exploiting the convexity of the gap function. □

Remark 7. This result again shows that it is beneficial to choose α_0 as large as possible, i.e. α_0 close to $1/3$ in order to reduce the constant on the right hand side. Similar to the case of overrelaxation, larger values of α_n leads to smaller primal step sizes τ and hence an inertial term might be less beneficial in presence of an explicit term ∇f.

Remark 8. Letting $\gamma = \tau L_f$ we find that the parameter α should satisfy

$$\alpha < \frac{\sqrt{16\gamma + 9} - 3}{2\gamma} - 1$$

in order for the left-hand side term in (26) to be positive (and then σ needs to be chosen accordingly). We point out that this condition is a bit more restrictive than the condition in [19]. This is due to the fact that our convergence proof is based on the Lipschitz continuity of ∇f rather than its co-coercivity, which leads to the loss of a factor 2 in the size of the primal step size τ relatively to the Lipschitz parameter L_f.

5 Acceleration for strongly convex problems

Here in this section, we slightly improve the results in [5] on accelerated algorithms. We address more precisely the natural generalization proposed in [9] (also [31]) and studied in [4] (where rates of convergence are proven). The main novelty with respect to [4] is a proof that in an ergodic sense, also the primal-dual gap is controlled and decreases at rate $O(1/N^2)$ where N is the number of iterations. In addition to our assumptions (i)-(iii) we assume that

(iv) f or g (or both) are strongly convex with respective parameters γ_f, γ_g and hence the primal objective is strongly convex with parameter $\gamma = \gamma_f + \gamma_g > 0$.

In fact, we observe that since

$$f(x) + g(x) = \left(f(x) - \frac{\gamma_f}{2}\|x\|^2\right) + \left(g(x) + \frac{\gamma_f}{2}\|x\|^2\right)$$

we can "transfer" the strong convexity of f to g: letting $\tilde{f} = f - \gamma_f\|\cdot\|^2/2$, $\tilde{g} = g + \gamma_f\|\cdot\|^2/2$, and $\gamma = \gamma_f + \gamma_g$, we have now that $\tilde{g}$ is γ-convex. In addition, $\nabla\tilde{f} = \nabla f - \gamma_f I$, so that

$$x' = (I + \tilde{\tau}\partial\tilde{g})^{-1}(x - \tilde{\tau}\nabla\tilde{f}(x)) \;\Leftrightarrow\; x' = (I + \tau\partial g)^{-1}(x - \tau\nabla f(x))$$

with

$$\tau = \frac{\tilde{\tau}}{1 + \gamma_f\tilde{\tau}}, \quad \text{so that } \tilde{\tau} := \frac{\tau}{1 - \gamma_f\tau} \tag{28}$$

(observe that τ needs, as expected, to be less than $1/\gamma_f > 1/L_f$). In addition, we find that $\nabla\tilde{f}$ is $(L_f - \gamma_f)$-Lipschitz. Hence in the following, to simplify we will just assume that g is strongly convex (that is, $\gamma_f = 0, \gamma = \gamma_g$), replacing assumption (iv) with the simpler assumption:

(iv') g is strongly convex with parameter $\gamma > 0$.

We must eventually mention here that in case $f = 0$, the dual problem, which has the form $\min_y g^*(-K^*y) + h^*(y)$, is the sum of a smooth plus a nonsmooth objective which could be tackled directly by more standard optimal methods [3, 22, 23] yielding similar convergence rates (provided one knows how to compute the Lipschitz gradient ∇g^*, which is slightly different from the assumptions we use in this paper).

5.1 Convergence analysis

With this additional assumption, the descent rule (9) can be slightly improved: indeed, thanks to the strong convexity of g, we can control an additional quadratic term on the right-hand side. It follows that for any $x \in \mathcal{X}$,

$$\begin{aligned} f(x) + g(x) + \tfrac{1}{\tau}D_x(x,\bar{x}) + \tfrac{L_f}{2}\|\hat{x} - \bar{x}\|^2 \geq \\ f(\hat{x}) + g(\hat{x}) + \langle K(\hat{x} - x), \tilde{y}\rangle + \tfrac{1}{\tau}D_x(\hat{x},\bar{x}) + \tfrac{1}{\tau}D_x(x,\hat{x}) + \tfrac{\gamma}{2}\|x - \hat{x}\|^2. \end{aligned} \tag{29}$$

It follows that (8) is also improved, with the additional term $\frac{\gamma}{2}\|x - \hat{x}\|^2$ on the left-hand side. One sees that one will be able to obtain a good convergence rate whenever the last two terms in (29) can be combined into one, which requires that $D_x(x,\hat{x}) = \frac{1}{2}\|x - \hat{x}\|_2^2$, that is, we must

consider linear proximity operators in the x variable.[1] To simplify the notation we will drop the subscript "2" in the norm for x, in the rest of this section.

Algorithm 4: $O(1/N^2)$ Accelerated primal-dual algorithm

- Input: Operator norm $L = \|K\|$, Lipschitz constant L_f of ∇f, parameter γ of strong convexity of g, and Bregman distance function D_y, $D_x(x, x') = \frac{1}{2}\|x - x'\|_x^2$
- Initialization: Choose $x^{-1} = x^0 \in \mathcal{X}$, $\tau_0, \sigma_0, \theta_0 > 0$ which satisfy (34).
- Iterations: For each $n \geq 0$ let

$$\begin{cases} (x^{n+1}, y^{n+1}) = \mathcal{PD}_{\tau_n,\sigma_n}(x^n, y^n, x^n + \theta_n(x^n - x^{n-1}), y^{n+1}) \\ \tau_{n+1}, \sigma_{n+1}, \theta_{n+1} \text{ satisfy (32), (33), (34).} \end{cases} \tag{30}$$

Now, we can specialize "à la" [5]. That is, we choose in (8) $\tilde{y} = \hat{y} = y^{n+1}$, $\hat{x} = x^{n+1}$, $\tilde{x} = x^n + \theta_n(x^n - x^{n-1})$, $\bar{x} = x^n$, $\bar{y} = y^n$, and make τ, σ depend also on the iteration counter n. In particular, now, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} &\langle K(x - \hat{x}), \hat{y} - \tilde{y}\rangle - \langle K(\hat{x} - \tilde{x}), y - \hat{y}\rangle \\ &\qquad = -\left\langle K(x^{n+1} - x^n), y - y^{n+1}\right\rangle + \theta_n \left\langle K(x^n - x^{n-1}), y - y^{n+1}\right\rangle \\ &\qquad = -\left\langle K(x^{n+1} - x^n), y - y^{n+1}\right\rangle + \theta_n \left\langle K(x^n - x^{n-1}), y - y^n\right\rangle \\ &\qquad\qquad + \theta_n \left\langle K(x^n - x^{n-1}), y^n - y^{n+1}\right\rangle . \end{aligned}$$

The last term is simply controlled by observing that

$$\begin{aligned} &\theta_n \left\langle K(x^n - x^{n-1}), y^n - y^{n+1}\right\rangle \\ &\qquad \geq -\theta_n L\|x^n - x^{n-1}\|\|y^n - y^{n+1}\| \geq -\theta_n L \left(\mu \frac{\|x^n - x^{n-1}\|^2}{2} + \frac{\|y^n - y^{n+1}\|^2}{2\mu}\right) \end{aligned}$$

for any $\mu > 0$, and choosing $\mu = \theta_n L \sigma_n$ we obtain

$$\begin{aligned} &\langle K(x - \hat{x}), \hat{y} - \tilde{y}\rangle - \langle K(\hat{x} - \tilde{x}), y - \hat{y}\rangle \\ &\qquad \geq -\left\langle K(x^{n+1} - x^n), y - y^{n+1}\right\rangle + \theta_n \left\langle K(x^n - x^{n-1}), y - y^n\right\rangle \\ &\qquad\qquad - \frac{\|y^{n+1} - y^n\|^2}{2\sigma_n} - (\theta_n^2 L^2 \sigma_n)\frac{\|x^n - x^{n-1}\|^2}{2} . \end{aligned}$$

To sum up, it follows from (8), with the additional strong convexity term from (29), that for any

[1] It must be observed here that the right assumption on g to obtain an accelerated scheme with an arbitrary Bregman distance D_x should be that g is "strongly convex with respect to ψ_x", in the sense that $g - \gamma\psi_x$ is convex. The proof would then be similar. However, it is not clear whether this covers very interesting situations beyond the standard case.

(x,y), using also that $D_y(y^{n+1},y^n)\geq \frac{1}{2}\|y^{n+1}-y^n\|^2$,

$$\begin{aligned}\frac{\|x-x^n\|^2}{2\tau_n}+\frac{D_y(y,y^n)}{\sigma_n}-\theta_n\left\langle K(x^n-x^{n-1}),y-y^n\right\rangle+\frac{\theta_n^2L^2\sigma_n}{2}\|x^n-x^{n-1}\|^2\\ \geq \mathcal{L}(x^{n+1},y)-\mathcal{L}(x,y^{n+1})+\frac{1+\gamma\tau_n}{2\tau_n}\|x-x^{n+1}\|^2\\ +\frac{D_y(y,y^{n+1})}{\sigma_n}-\left\langle K(x^{n+1}-x^n),y-y^{n+1}\right\rangle+\frac{1-L_f\tau_n}{2\tau_n}\|x^{n+1}-x^n\|^2. \quad (31)\end{aligned}$$

Assume the sequences θ_n,τ_n,σ_n satisfy for all $n\geq 0$

$$\frac{1+\gamma\tau_n}{\tau_n}\geq\frac{1}{\theta_{n+1}\tau_{n+1}}, \quad (32)$$

$$\frac{1}{\sigma_n}=\frac{1}{\theta_{n+1}\sigma_{n+1}}, \quad (33)$$

$$L^2\sigma_n\leq\frac{1}{\tau_n}-L_f. \quad (34)$$

Then (31) becomes (using $\theta_n^2L^2\sigma_n=\theta_nL^2\sigma_{n-1}$, thanks to (33))

$$\begin{aligned}\frac{\|x-x^n\|^2}{2\tau_n}+\frac{D_y(y,y^n)}{2\sigma_n}+\theta_n\left(L^2\sigma_{n-1}\frac{\|x^n-x^{n-1}\|^2}{2}-\left\langle K(x^n-x^{n-1}),y-y^n\right\rangle\right)\\ \geq \mathcal{L}(x^{n+1},y)-\mathcal{L}(x,y^{n+1})+\frac{1}{\theta_{n+1}}\left(\frac{\|x-x^{n+1}\|^2}{2\tau_{n+1}}+\frac{D_y(y,y^{n+1})}{2\sigma_{n+1}}\right.\\ \left.+\theta_{n+1}\left(L^2\sigma_n\frac{\|x^{n+1}-x^n\|^2}{2}-\left\langle K(x^{n+1}-x^n),y-y^{n+1}\right\rangle\right)\right). \quad (35)\end{aligned}$$

Observe that from (33), $\prod_{n=1}^N\theta_n=\sigma_0/\sigma_N$. We now let

$$T_N=\sum_{n=1}^N\frac{\sigma_{n-1}}{\sigma_0},\quad X^N=\frac{1}{T_N}\sum_{n=1}^N\frac{\sigma_{n-1}}{\sigma_0}x^n,\quad Y^N=\frac{1}{T_N}\sum_{n=1}^N\frac{\sigma_{n-1}}{\sigma_0}y^n. \quad (36)$$

Then, summing (35) (first multiplied on both sides by σ_n/σ_0) from $n=0$ to $n=N-1$ and assuming $x^{-1}=x^0$, using also the convexity of $(\xi,\eta)\mapsto\mathcal{L}(\xi,y)-\mathcal{L}(x,\eta)$ (for fixed x,y), we deduce

$$\begin{aligned}T_N\left(\mathcal{L}(X^N,y)-\mathcal{L}(x,Y^N)\right)+\frac{\sigma_N}{\sigma_0}\left(\frac{\|x-x^N\|^2}{2\tau_N}+\frac{D_y(y,y^N)}{\sigma_N}\right.\\ \left.+\theta_N\left(L^2\sigma_{N-1}\frac{\|x^N-x^{N-1}\|^2}{2}-\left\langle K(x^N-x^{N-1}),y-y^N\right\rangle\right)\right)\\ \leq\frac{\|x-x^0\|^2}{2\tau_0}+\frac{D_y(y,y^0)}{\sigma_0}.\end{aligned}$$

Considering eventually that (using again (33))

$$\left\langle K(x^N-x^{N-1}),y-y^N\right\rangle\leq\frac{D_y(y,y^N)}{\theta_N\sigma_N}+\frac{L^2\sigma_{N-1}}{2}\|x^N-x^{N-1}\|^2,$$

we deduce the following result.

Theorem 4. *Assume $D_x(x,x') = \frac{1}{2}\|x-x'\|_x^2$. Let (x^n,y^n) be a sequence generated by Algorithm 4, and let (X^N,Y^N) and (T_N) be the ergodic averages given by (36). Then, for all x and y, for all $N \geq 1$, one has the estimate*

$$T_N\left(\mathcal{L}(X^N,y)-\mathcal{L}(x,Y^N)\right)+\frac{\sigma_N}{\sigma_0}\frac{\|x-x^N\|^2}{2\tau_N} \leq \frac{\|x-x^0\|^2}{2\tau_0}+\frac{D_y(y,y^0)}{\sigma_0}. \tag{37}$$

Remark 9. Notice that, taking $(x,y)=(x^*,y^*)$ a saddle-point in (37), we find that X^N and x^N are bounded (and converge to x^*). If we assume that h has full domain, so that $h^*(y)/|y| \to \infty$ as $|y| \to \infty$, we deduce that also Y^N is bounded (since otherwise $-\mathcal{L}(x^*,Y^N)$ would go to infinity), and it follows that the (x,y) which realize the supremum in the expression $\mathcal{L}(X^N,y)-\mathcal{L}(x,Y^N)$ are also globally bounded. It follows the global estimate on the gap

$$\sup_{x,y}\mathcal{L}(X^N,y)-\mathcal{L}(x,Y^N) \leq \frac{C}{T_N}. \tag{38}$$

5.2 Parameter choices

It turns out that it is possible to choose sequences $(\tau_n,\sigma_n,\theta_n)$ satisfying (32), (33), (34) in order to have $1/T_N = O(1/N^2)$. A possible choice, similar to [5], to ensure (32), (33), (34) is to keep the product $\sigma_n\tau_n$ constant and let

$$\theta_{n+1}=\frac{1}{\sqrt{1+\gamma\tau_n}}, \quad \tau_{n+1}=\theta_{n+1}\tau_n\ , \quad \sigma_{n+1}=\sigma_n/\theta_{n+1}. \tag{39}$$

Then, letting

$$\tau_0=\frac{1}{2L_f}\ , \quad \sigma_0=\frac{L_f}{L^2}$$

(or $\tau_0=\sigma_0=1/L$ if $L_f=0$), we find that by induction, since $\tau_{n+1}/\tau_n=\sigma_n/\sigma_{n+1}<1$ for each n, (34) will be satisfied. We refer to [5] for a proof that this choice ensures that $\sigma_n \sim \gamma n/(4L^2)$, so that $T_N \sim \gamma N^2/(4L_f)$.

A simpler (still slightly suboptimal) choice is to take $\sigma_0>0$ arbitrary, and

$$\tau_n=\frac{2}{\gamma n+2(L_f+L^2\sigma_0)}, \quad \sigma_n=\sigma_0+\frac{\gamma n\sigma_0}{\gamma+2(L_f+L^2\sigma_0)}. \tag{40}$$

Then, (32), (33), (34) hold, and

$$T_N=N+\frac{N(N-1)}{2}\frac{\gamma}{\gamma+2(L_f+L^2\sigma_0)}. \tag{41}$$

Observe that in this case,

$$\theta_{n+1}=\frac{\sigma_n}{\sigma_{n+1}}=\frac{\gamma(n+1)+2(L_f+L^2\sigma_0)}{\gamma(n+2)+2(L_f+L^2\sigma_0)}$$

and

$$\theta_{n+1}\tau_{n+1}=\frac{2}{\gamma(n+2)+2(L_f+L^2\sigma_0)}=\frac{\tau_n}{1+\gamma\tau_n},$$

that is, the equality holds in (32).

The optimal rule should consist in choosing equalities in (32), (33) and (34). We find that $\sigma_0 > 0$ can be chosen arbitrarily,

$$\tau_0 = \frac{1}{L_f + L^2\sigma_0},$$

and then:

$$1 + \gamma\tau_n = \frac{\tau_n}{\tau_{n+1}}\frac{\sigma_{n+1}}{\sigma_n} = \frac{\tau_n^2}{\tau_{n+1}^2}\frac{1 - L_f\tau_{n+1}}{1 - L_f\tau_n},$$

$$\frac{\tau_{n+1}^2}{1 - L_f\tau_{n+1}} = \frac{\tau_n^2}{(1 + \gamma\tau_n)(1 - L_f\tau_n)} =: \beta_{n+1}^2$$

so that, assuming $L_f\tau_n < 1$ (which is true for $n = 0$),

$$\begin{aligned}\tau_{n+1} &= \beta_{n+1}\Big(\sqrt{1 + \tfrac{L_f^2}{4}\beta_{n+1}^2} - \frac{L_f}{2}\beta_{n+1}\Big) \\ &= \frac{\beta_{n+1}}{\sqrt{1 + \frac{L_f^2}{4}\beta_{n+1}^2} + \frac{L_f}{2}\beta_{n+1}} = \frac{\tau_n}{\sqrt{(1 + \gamma\tau_n)(1 - L_f\tau_n) + \frac{L_f^2}{4}\tau_n^2} + \frac{L_f}{2}\tau_n} \leq \tau_n,\end{aligned}$$

and

$$\theta_{n+1} = \frac{\sqrt{(1 + \gamma\tau_n)(1 - L_f\tau_n) + \frac{L_f^2}{4}\tau_n^2} + \frac{L_f}{2}\tau_n}{1 + \gamma\tau_n} \in \left(\frac{1}{1 + \gamma\tau_n}, \frac{1}{\sqrt{1 + \gamma\tau_n}}\right).$$

Let us denote τ_n^{opt}, σ_n^{opt} and T_N^{opt} the τ_n, σ_n and T_N obtained by this "optimal" rule (and the corresponding T_N) and let us keep the notation τ_n, σ_n, T_N for the expressions in (40) and (41). These choices, in particular, satisfy the equality in (32), (33), but a strict inequality (for $n \geq 1$) in (34). We assume that the starting point $\sigma_0 = \sigma_0^{opt}$ is the same, then of course also $\tau_0 = \tau_0^{opt}$. Then one has:

Lemma 2. *For each $n \geq 0$, $\sigma_n^{opt} \geq \sigma_n$, and $T_n^{opt} \geq T_n \sim cn^2$.*

Proof. We proceed by induction. We assume $\sigma_n^{opt} \geq \sigma_n$, which is true for $n = 0$. For practical reasons, let us set $X_n^{opt} = L^2\sigma_n^{opt}$, $Y_n^{opt} = -1/\tau_n^{opt}$, $X_n = L^2\sigma_n$, and $Y_n = -1/\tau_n$. Then from the equality in (32), we have for all n

$$X_{n+1}Y_{n+1} = X_nY_n - \gamma X_n, \quad X_{n+1}^{opt}Y_{n+1}^{opt} = X_n^{opt}Y_n^{opt} - \gamma X_n^{opt}, \tag{42}$$

We also assume $\Pi_n := X_nY_n \geq \Pi_n^{opt} := X_n^{opt}Y_n^{opt}$, which is true at $n = 0$. It follows then that from (42) and $X_n^{opt} \geq X_n$ that $\Pi_{n+1} \geq \Pi_{n+1}^{opt}$. Observe that being this product negative, it means in fact that $|\Pi_{n+1}| \leq |\Pi_{n+1}^{opt}|$.

On the other hand, from (34), one has that

$$\Sigma_{n+1} := X_{n+1} + Y_{n+1} \leq -L_f = X_{n+1}^{opt} + Y_{n+1}^{opt} =: \Sigma_{n+1}^{opt} \leq 0$$

(and, again, $|\Sigma_{n+1}| \geq |\Sigma_{n+1}^{opt}|$).

One has then

$$X_{n+1} = \frac{\Sigma_{n+1} + \sqrt{\Sigma_{n+1}^2 - 4\Pi_{n+1}}}{2} = \frac{\sqrt{\Sigma_{n+1}^2 + 4|\Pi_{n+1}|} - \sqrt{\Sigma_{n+1}^2}}{2},$$

which, by concavity of $\sqrt{\cdot}$ and since $\Sigma_{n+1}^2 \geq (\Sigma_{n+1}^{opt})^2$, $|\Pi_{n+1}| \leq |\Pi_{n+1}^{opt}|$, is less than the similar expression for X_{n+1}^{opt}. This shows the Lemma. □

6 Acceleration for smooth and strongly convex problems

In this section, we finally make the additional assumption that

(v) h^* is strongly convex with parameter $\delta > 0$.

Equivalently, h has $(1/\delta)$-Lipschitz gradient, so that the primal objective is both smooth and strongly convex. Then, as expected, the rate can be improved, to linear convergence. In this section, we must assume that the Bregman distance functions satisfy $D_x(x, x') = \frac{1}{2}\|x - x'\|_2^2$ and $D_y(y, y') = \frac{1}{2}\|y - y'\|_2^2$, that is, we are for both variables in the Euclidean/Hilbertian setting. For simplicity we will drop the subscript "2" in the rest of this section.

We show here how to modify the proof of the previous case to obtain a linear convergence rate on the gap. This slightly improves the results in [5, 4] which only give a rate on the iterates. In contrast to [5], we do not show here convergence for a large range of relaxation parameters θ, but the proof presented can handle the explicit term ∇f and yields a similar convergence rate.

6.1 Convergence analysis

Algorithm 5: $O(\theta^N)$ Accelerated primal-dual algorithm

- Input: Operator norm $L = \|K\|$, Lipschitz constant L_f of ∇f, parameters γ, δ of strong convexity of g and h^*, $D_x(x, x') = \frac{1}{2}\|x - x'\|_x^2$ and $D_y(y, y') = \frac{1}{2}\|y - y'\|_y^2$.
- Initialization: Choose $x^{-1} = x^0 \in \mathcal{X}$, $\tau, \sigma, \theta > 0$ which satisfy (44) and (45).
- Iterations: For each $n \geq 0$ let

$$(x^{n+1}, y^{n+1}) = \mathcal{PD}_{\tau,\sigma}(x^n, y^n, x^n + \theta(x^n - x^{n-1}), y^{n+1}) \tag{43}$$

A first remark is that the inequality (31), in case h^* is δ-convex, can be written

$$\begin{aligned}\frac{\|x - x^n\|^2}{2\tau} + \frac{\|y - y^n\|^2}{2\sigma} - \theta\left\langle K(x^n - x^{n-1}), y - y^n\right\rangle + \frac{\theta^2 L^2 \sigma}{2}\|x^n - x^{n-1}\|^2 \\ \geq \mathcal{L}(x^{n+1}, y) - \mathcal{L}(x, y^{n+1}) + \frac{1 + \gamma\tau}{2\tau}\|x - x^{n+1}\|^2 \\ + \frac{1 + \delta\sigma}{2\sigma}\|y - y^{n+1}\|^2 - \left\langle K(x^{n+1} - x^n), y - y^{n+1}\right\rangle + \frac{1 - L_f\tau}{2\tau}\|x^{n+1} - x^n\|^2.\end{aligned}$$

It follows that if one can choose τ, σ, θ so that

$$1 + \gamma\tau = 1 + \delta\sigma = \frac{1}{\theta} \tag{44}$$

$$\frac{1 - L_f\tau}{\tau} \geq \theta L^2 \sigma \tag{45}$$

then, multiplying the inequality by θ^{-n} and summing from $n = 0$ to $N - 1$, we get (assuming

$x^{-1} = x^0$)

$$\frac{\|x - x^0\|^2}{2\tau} + \frac{\|y - y^0\|^2}{2\sigma} \geq \sum_{n=1}^{N} \frac{1}{\theta^{n-1}} (\mathcal{L}(x^n, y) - \mathcal{L}(x, y^n)) + \frac{1}{\theta^N} \left(\frac{\|x - x^N\|^2}{2\tau} + \frac{\|y - y^N\|^2}{2\sigma} - \theta \left\langle K(x^N - x^{N-1}), y - y^N \right\rangle \right) + \frac{1 - L_f \tau}{2\tau \theta^{N-1}} \|x^N - x^{N-1}\|^2.$$

Using (45) again, we deduce

$$\sum_{n=1}^{N} \frac{1}{\theta^{n-1}} (\mathcal{L}(x^n, y) - \mathcal{L}(x, y^n)) + \frac{\|x - x^N\|^2}{2\tau\theta^N} \leq \frac{\|x - x^0\|^2}{2\tau} + \frac{\|y - y^0\|^2}{2\sigma}.$$

Hence, letting now

$$T_N = \sum_{n=1}^{N} \theta^{-n+1} = \frac{1 - \theta^N}{1 - \theta} \frac{1}{\theta^{N-1}} \quad \text{and} \quad Z^N = (X^N, Y^N) = \frac{1}{T_N} \sum_{n=1}^{N} \theta^{-n+1} z^n \tag{46}$$

we obtain the following result

Theorem 5. *Assume $D_x(x, x') = \frac{1}{2}\|x - x'\|_x^2$ and $D_y(y, y') = \frac{1}{2}\|y - y'\|_y^2$. Let (x^n, y^n) be a sequence generated by Algorithm 5, and let (X^N, Y^N) and (T_N) be the ergodic averages defined in (46). Then, for all x and y, for all $N \geq 1$, one has the estimate*

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) + \frac{\theta(1 - \theta)}{1 - \theta^N} \frac{\|x - x^N\|^2}{2\tau} \leq \frac{1}{T_N} \left(\frac{\|x - x^0\|^2}{2\tau} + \frac{\|y - y^0\|^2}{2\sigma} \right) \tag{47}$$

which yields a linear convergence rate.

6.2 Parameter choices

Solving the equations (44) for τ, σ, θ, we obtain, letting [2]

$$\mu = \frac{\delta(\gamma + L_f)}{2L^2} \left(\sqrt{1 + 4\frac{\gamma L^2}{\delta(\gamma + L_f)^2}} - 1 \right) \in (0, 1) :$$

$$\tau = \frac{\mu}{\gamma(1 - \mu)}, \quad \sigma = \frac{\mu}{\delta(1 - \mu)}, \quad \theta = 1 - \mu. \tag{48}$$

In case $L_f = 0$, one has

$$\mu = \frac{\delta\gamma}{2L^2} \left(\sqrt{1 + 4\frac{L^2}{\delta\gamma}} - 1 \right)$$

and the above formulas simplify to

$$\tau = \delta \frac{1 + \sqrt{1 + 4\frac{L^2}{\gamma\delta}}}{2L^2}, \quad \sigma = \gamma \frac{1 + \sqrt{1 + 4\frac{L^2}{\gamma\delta}}}{2L^2}, \quad \theta = 1 - \frac{\delta\gamma}{2L^2} \left(\sqrt{1 + 4\frac{L^2}{\delta\gamma}} - 1 \right). \tag{49}$$

[2] using WolframAlpha to check our calculations.

In this case, if $\delta\gamma \ll L^2$, then the linear rate of convergence is governed by the factor

$$\theta \leq 1 - \frac{\sqrt{\delta\gamma}}{L} + \frac{\delta\gamma}{2L^2} \approx 1 - \frac{\sqrt{\delta\gamma}}{L} :$$

we remark that this is of the same order as the rate found in [5] (which was not a bound on the gap though). Note however that a more refined analysis in [5] allowed to obtain better rates, for a larger choice of parameters θ. A similar analysis with $L_f \neq 0$ would probably lead to close results through very tedious calculations. On the other hand, the new proof allows for slightly larger steps than in [5].

7 Computational examples

In this section we demonstrate the practical performance of the proposed algorithms on a number of randomly generated instances of classical optimization problems.

7.1 Matrix games

Here, we consider the following min-max matrix game:

$$\min_{x\in\Delta_l} \max_{y\in\Delta_k} \mathcal{L}(x,y) = \langle Kx, y\rangle\,, \tag{50}$$

where Δ_k and Δ_l denote the standard unit simplices in $\mathbb{R}^k$ and $\mathbb{R}^l$ and $K \in \mathbb{R}^{k\times l}$. This class of min-max matrix games can be used for approximately finding the Nash equilibrium of two-person zero-sum matrix games such as two-person Texas Hold'em Poker. Following the computational experiments in [23], the entries of K are independently and uniformly distributed in the intervall $[-1, 1]$. We denote by $L = \|K\|$ the operator norm of K. We can also easily compute the primal-dual gap of a feasible pair (x, y). For this we observe that $\arg\min_{x\in\Delta_l} \mathcal{L}(x, y) = e_j$, where $e_j \in \Delta_l$ is the j-th standard basis vector corresponding to the smallest entry of the vector $K^T y$. Likewise, $\arg\max_{y\in\Delta_k} \mathcal{L}(x, y) = e_i$, where i corresponds to the coordinate of the largest entry of Kx. Hence, the primal-dual gap is computed as

$$\mathcal{G}(x,y) = \left[\mathcal{P}(x) = \max_i (Kx)_i\right] - \left[\mathcal{D}(y) = \min_j (K^T y)_j\right]$$

7.1.1 Linear and nonlinear primal-dual algorithms

We first consider different Bregman distance settings of the nonlinear primal-dual algorithm presented in Algorithm 1. The initial points (x^0, y^0) are chosen to be the centers of the simplices, that is $x_j^0 = 1/l$ and $y_i^0 = 1/k$ for all i, j. There are two obvious choices for the Bregman distance functions:

1. **Euclidean setting:** In the Euclidean setting, $\|\cdot\|_x = \|\cdot\|_y = \|\cdot\|_2$, $D_x(x, x') = \frac{1}{2}\|x - x'\|^2$, and $D_y(y, y') = \frac{1}{2}\|y - y'\|^2$. It follows that $\max_{x\in\Delta_l} D_x(x, x^0) = (1 - \frac{1}{l})/2$ and likewise $\max_{y\in\Delta_k} D_y(y, y^0) = (1 - \frac{1}{k})/2$. The operator norm is computed as the largest singular

value $L_2 = s_{\max}(K)$. In the iterates of the algorithm, we need to solve subproblems of the following form:

$$\hat{x} = \arg\min_{x \in \Delta_l} \langle x, \xi \rangle + \frac{\|x - \bar{x}\|^2}{2\tau} \Leftrightarrow \hat{x} = \mathrm{proj}_{\Delta_l}(\bar{x} - \tau\xi),$$

where we are using the $n \log n$ algorithm described in [12] to compute the orthogonal projections on the unit simplices. Taking the supremum on the right hand side of (17), the ergodic $O(1/N)$ rate for the primal-dual gap bounded by

$$\mathcal{G}(X^N, Y^N) \leq \frac{2}{N}\left(\frac{1 - \frac{1}{l}}{2\tau} + \frac{1 - \frac{1}{k}}{2\sigma}\right).$$

Since $\tau\sigma L_2^2 = 1$, the right hand side is minimized by setting

$$\tau = \frac{1}{L_2}\sqrt{\frac{1 - \frac{1}{l}}{1 - \frac{1}{k}}}, \quad \sigma = \frac{1}{L_2}\sqrt{\frac{1 - \frac{1}{k}}{1 - \frac{1}{l}}}.$$

Hence we get the following final estimate for the gap

$$\mathcal{G}(X^N, Y^N) \leq \frac{2\sqrt{(1 - \frac{1}{l})(1 - \frac{1}{k})}}{N} L_2.$$

2. **Entropy setting:** In the entropy setting we choose $\|\cdot\|_x = \|\cdot\|_y = \|\cdot\|_1$ and the Bregman distance functions are given by $D_x(x, x') = \sum_j x_j(\log x_j - \log x'_j) - x_j + x'_j$ and $D_y(y, y') = \sum_i y_i(\log y_i - \log y'_i) - y_i + y'_i$, which are 1-convex with respect to the 1-norm. Now, $\max_{x \in \Delta_l} D_x(x, x^0) = \log l$ and $\max_{y \in \Delta_k} D_y(y, y^0) = \log k$. The operator norm is given by $L_1 = \sup_{\|x\|_1 = 1} \|Kx\|_\infty = \max_{i,j} |K_{i,j}|$.

   It is well known that in the entropy setting, the iterates of the algorithm are explicit:

$$\hat{x} = \arg\min_{x \in \Delta_l} \langle x, \xi \rangle + \frac{1}{\tau} D_x(x, \bar{x}) \Leftrightarrow \hat{x}_j = \frac{\bar{x}_j \exp(-\tau\xi_j)}{\sum_{j=1}^l \bar{x}_j \exp(-\tau\xi_j)}$$

   In turn, the ergodic $O(1/N)$ rate in (17) specializes to

$$\mathcal{G}(X^N, Y^N) \leq \frac{2}{N}\left(\frac{\log l}{\tau} + \frac{\log k}{\sigma}\right).$$

   Again, the right hand side is minimized by choosing

$$\tau = \frac{1}{L_1}\sqrt{\frac{\log l}{\log k}}, \quad \sigma = \frac{1}{L_1}\sqrt{\frac{\log k}{\log l}}$$

   We obtain the final estimate as

$$\mathcal{G}(X^N, Y^N) \leq \frac{4\sqrt{\log l \log k}}{N} L_1.$$

Remark 10. Since the entries of K are uniformly distributed in $[-1,1]$, the solutions of the min-max matrix games always cluster around the simplex centers $1/k$ and $1/l$. In this region, the entropy functions are strongly k- and l-convex, respectively, and hence for small variations around the center,

$$D_x(x,x') \gtrapprox \frac{1}{2}(\sqrt{l}\|x-x'\|_2)^2, \quad D_y(y,y') \gtrapprox \frac{1}{2}(\sqrt{k}\|y-y'\|_2)^2.$$

With this information we can get a better (local) estimate of the operator norm:

$$L_{\text{cent}} \approx \sup_{\|x\|_2 \leq 1/\sqrt{l}} \|Kx\|_2/\sqrt{k} = L_2/\sqrt{kl}$$

In practice, we observed that slightly larger values, e.g. $1.7 \cdot L_{\text{cent}}$ worked very well in our experiments. It would be of great interest to find a method which is able to exploit this variability, unfortunately we were not able to find a rigorous and efficient method.

First let us observe that our theoretical worst case bounds for the matrix games are exactly the same as the corresponding worst case bounds in [23]. In Table 1 we report the number of iterations of the $O(1/N)$ primal-dual algorithms using the Euclidean setting and the entropy setting to reach a primal-dual gap that is less than ε. One can see that the entropy-based algorithm is consistently faster compared to the Euclidean-based algorithm. Furthermore, one can see that the complexity for finding an ε accurate solution grows, as predicted in Theorem 1, with a factor of order $1/\varepsilon$. Indeed, one can see that reducing ε by a factor of 10 roughly leads to 10 times more iterations. Comparing the results with the results reported in [23] the proposed algorithms are significantly faster. Also observe that counterintuively, less iterations are needed for larger problems. This might be due to the fact that the value of the gap of theses problems at the centers of the simplices goes to zero as the size goes to infinity, making this initialization more beneficial for larger problems.

Table 1: Computational results of Algorithm 1 applied to the matrix game problem (50).

	$\varepsilon = 10^{-3}$		$\varepsilon = 10^{-4}$	
k/l	Euclidean	Entropy	Euclidean	Entropy
100/100	942	730	9394	7292
100/500	760	750	7671	7378
100/1000	1138	960	11330	9862
500/100	1085	648	10743	6474
500/500	483	333	4782	3290
500/1000	480	350	4796	3430
1000/100	1537	640	15394	6284
1000/500	547	297	5434	2905
1000/1000	381	261	3797	2546

7.1.2 Ergodic versus nonergodic sequence

We also tested the performance of the nonergodic sequences, i.e. the rate of convergence of the primal-dual gap of the iterates (x^n, y^n). Figure 1 depicts logarithmic convergence plots in the setting $k = l = 1000$, for both the Euclidean and the entropy setting. It shows that in the Euclidean setting, the nonergodic sequence converges even faster than the ergodic sequence. In the entropy setting however, we observed the contrary. The nonergodic sequence converges much slower than the ergodic sequence. We do not know the reason for this behavior. For both ergodic sequences, the gap decreases exactly at rate $O(1/N)$ as predicted by the analysis.

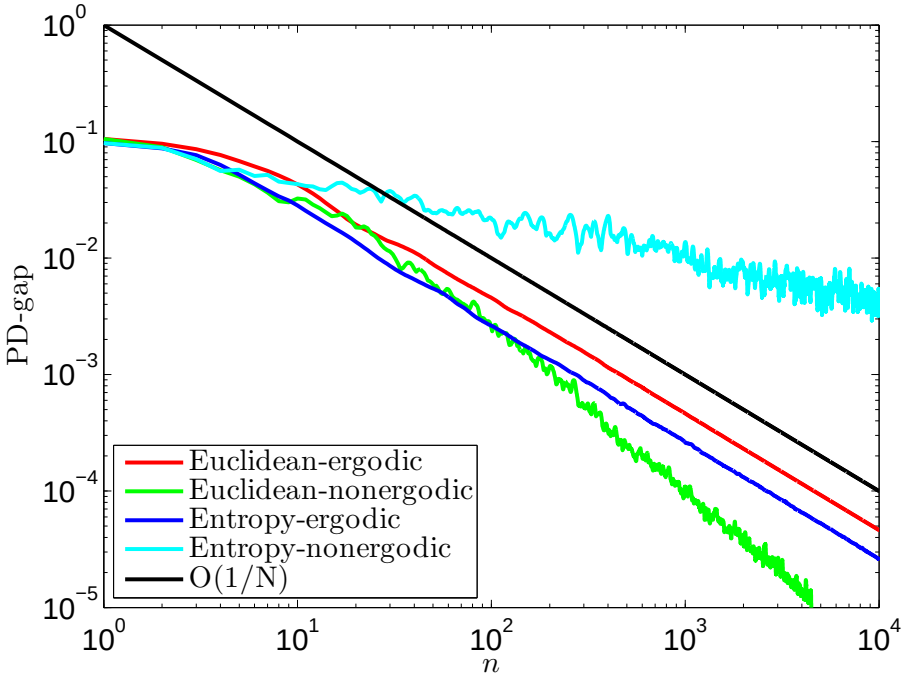

Figure 1: Comparison between the performance of the ergodic and the nonergodic sequences of Algorithm 1 applied to the matrix game problem (50).

7.1.3 Overrelaxed and inertial primal-dual algorithms

In this section, we evaluate the performance of the overrelaxed and inertial version of the Euclidean primal-dual algorithm detailed in Algorithm 2 and Algorithm 3. We vary the relaxation parameter ρ and the inertial parameter α (which are kept constant during the iterations) and record the number of iterations that are necessary to reach a primal-dual gap which is less a tolerance of $\varepsilon = 10^{-4}$. For both, the inertial and overrelaxed versions, we observe that the algorithms are still converging for the theoretical limits $\rho = 2$ and $\alpha = 1/3$.

In Table 2, we report the number of iterations using different values of the relaxation param-

eter ρ. As predicted in Theorem 2, the number of iterations are approximately proportional to the factor $1/\rho$. In Table 3, we report the number of iterations using different inertial parameters α. Again, as predicted in Theorem 3, the number of iterations roughly correspond to the factor $1 - 1/\alpha$.

Table 2: Computational results of Algorithm 2 applied to the matrix game problem (50).

k/l	$\rho = 1$	$\rho = 5/4$	$\rho = 3/2$	$\rho = 7/4$	$\rho = 2$
100/100	9394	7224	5784	4825	4075
100/500	7671	5853	4748	3983	3337
100/1000	11330	8861	7225	6063	5151
500/100	10743	8389	6818	5721	4691
500/500	4782	3651	2924	2402	2051
500/1000	4796	3809	3156	2696	2346
1000/100	15394	11982	9721	8126	6835
1000/500	5434	4224	3426	2860	2437
1000/1000	3797	2975	2433	2050	1881

Table 3: Computational results of Algorithm 3 applied to matrix game problem (50).

k/l	$\alpha = 0$	$\alpha = 1/12$	$\alpha = 1/6$	$\alpha = 1/4$	$\alpha = 1/3$
100/100	9394	8660	7939	7234	6545
100/500	7671	7065	6467	5882	5328
100/1000	11330	10420	9520	8632	7751
500/100	10743	9882	9034	8203	7411
500/500	4782	4401	4026	3655	3291
500/1000	4796	4410	4029	3656	3299
1000/100	15394	14147	12912	11692	10514
1000/500	5434	4996	4563	4134	3721
1000/1000	3797	3492	3191	2897	2613

7.2 Simplex constrained least squares problem

In this section, we consider the following class of simplex-constrained least squares problems

$$\min_{x \in \Delta_l} \mathcal{P}(x) = \frac{1}{2}\|Kx - b\|^2, \tag{51}$$

where Δ_l again denotes the standard unit simplex in $\mathbb{R}^l$ and $K \in \mathbb{R}^{k \times l}$, and $b \in \mathbb{R}^k$. Several standard optimization problems used in machine learning such as the support vector machine can be obtained as special cases from (51). Here, K and b are randomly generated with their entries

uniformly and independently distributed in the intervall $[-1, 1]$. We again denote by $L = \|K\|$ the operator norm of K. The saddle-point formulation of (51) is given by

$$\min_{x\in\Delta_l} \max_y \mathcal{L}(x, y) = \langle Kx, y\rangle - b^T y - \frac{1}{2}\|y\|^2. \tag{52}$$

Furthermore, the dual problem is given by

$$\max_y \mathcal{D}(y) = \min_j (K^T y)_j - b^T y - \frac{1}{2}\|y\|^2$$

In turn, the primal-dual gap for a pair (x, y) can be easily computed by observing that $\arg\min_{x\in\Delta_l} \mathcal{L}(x, y) = e_j$ and also $\arg\max_y \mathcal{L}(x, y) = Kx - b$:

$$\mathcal{G}(x, y) = \left[\frac{1}{2}\|Kx - b\|^2\right] - \left[\min_j (K^T y)_j - b^T y - \frac{1}{2}\|y\|^2\right]$$

7.2.1 Accelerated primal-dual algorithms

Note that since the saddle-point problem is strongly convex in y, we can use the accelerated primal-dual algorithm presented in Algorithm 4 (by interchanging the role of the primal and the dual variables). Since $L_f = 0$, we apply the simple parameter choice (39). We initialize the algorithms with the obvious choice $(x^0)_j = 1/l$ for all j and $y^0 = Kx^0 - b$. Recall that in the accelerated algorithm, we have fixed $\|\cdot\|_y = \|\cdot\|_2$ and hence $D_y(y, y') = \frac{1}{2}\|y - y'\|^2$. Let us now consider two different setups of the algorithm:

1. **Euclidean setting:** In the Euclidean setting, we set $\|\cdot\|_x = \|\cdot\|_2$ and hence $D_x(x, x') = \frac{1}{2}\|x - x'\|^2$. The operator norm $L_2 = \|K\|$ is again given by $s_{\max}(K)$. According to (37), we guaranteed that after N iterations for all (x, y) it holds that

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) \leq \frac{\|x - x^0\|^2}{2\tau_0} + \frac{\|y - y^0\|^2}{2\sigma_0}$$

Substituting $y = \arg\max_y \mathcal{L}(X^N, y) = KX^N - b$, we obtain for all x

$$T_N(\mathcal{P}(X^N) - \mathcal{L}(x, Y^N)) \leq \frac{\|x - x^0\|^2}{2\tau_0} + \frac{\|K(X^N - x^0)\|^2}{2\sigma_0}$$

Taking the supremum with respect to x on both sides, it follows

$$\begin{aligned} T_N\mathcal{G}(X^N, Y^N) &\leq \sup_{x\in\Delta_l} \frac{\|x - x^0\|^2}{2\tau_0} + \frac{\|K(X^N - x^0)\|^2}{2\sigma_0} \\ &\leq \frac{1 - \frac{1}{l}}{2\tau_0} + \frac{\|K(X^N - x^0)\|^2}{2\sigma_0} \\ &\leq \frac{1 - \frac{1}{l}}{2\tau_0} + \frac{(1 - \frac{1}{l})}{2\sigma_0} L_2^2. \end{aligned}$$

The right hand side is minimized by choosing $\tau_0 = 1/L_2^2$ and $\sigma_0 = 1$ which gives the final estimate

$$\mathcal{G}(X^N, Y^N) \leq \frac{1 - \frac{1}{l}}{T_N} L_2^2,$$

where $T_N \sim O(N^2)$ is defined in (36).

2. **Entropy setting:** In the entropy setting, we choose $\|\cdot\|_x = \|\cdot\|_1$ and $D_x(x, x') = \sum_j x_j(\log x_j - \log x'_j) - x_j + x'_j$. The operator norm is now given by $L_{12} = \|K\| = \sup_{\|x\|_1 \leq 1} \|Kx\|_2 = \max_j \|K_j\|_2$ where K_j denotes the j-th column of K, which is typically smaller than L_2. In analogy to the above calculations, we have

$$\begin{aligned} T_N \mathcal{G}(X^N, Y^N) \leq & \sup_x \frac{D_x(x, x^0)}{\tau_0} + \frac{\|K(X^N - x^0)\|^2}{2\sigma_0} \\ \leq & \frac{\log l}{\tau_0} + \frac{L_2^2(1 - \frac{1}{l})}{2\sigma_0}. \end{aligned}$$

The optimal choice for τ_0 and σ_0 is now

$$\tau_0 = \sqrt{\frac{2 \log l}{L_{12}^2 L_2^2 (1 - \frac{1}{l})}}, \quad \sigma_0 = \sqrt{\frac{L_2^2(1 - \frac{1}{l})}{2 L_{12}^2 \log l}},$$

which yields the final estimate

$$\mathcal{G}(X^N, Y^N) \leq \frac{L_{12} L_2 \sqrt{\frac{(1 - \frac{1}{l}) \log l}{2}}}{T_N}.$$

We also observed that in the entropy setting, we can choose larger step sizes: choosing $L_{12} = 0.35 \cdot \max_j \|K_j\|_2$ gave experimentally good results. In Table 4, we report the number of iterations for Algorithm 4 in the Euclidean and the entropy setting. One can see that in the entropy setting, the algorithm converges significantly faster. Furthermore, one can see that the number of iterations which are necessary to reach a primal-dual gap less than ε nicely reflect the $O(1/N^2)$ rate of Algorithm 4. Indeed, reducing ε by a factor of 10 roughly leads to $\sqrt{10} \approx 3.16$ more iterations.

Table 4: Computational results of Algorithm 4 applied to the simplex constrained least squares problem (51).

	$\varepsilon = 10^{-3}$		$\varepsilon = 10^{-4}$	
k/l	Euclidean	Entropy	Euclidean	Entropy
100/100	423	128	1264	396
100/500	645	179	1881	563
100/1000	1008	191	2946	600
500/100	1039	243	3187	726
500/500	1399	329	4276	1026
500/1000	1530	365	4570	1142
1000/100	1752	367	5508	1115
1000/500	2257	459	7079	1425
1000/1000	2418	499	7507	1554

7.2.2 Ergodic versus nonergodic sequence

We also investigated the performance difference between the ergodic and the nonergodic sequences. Figure 2 shows a comparison between the ergodic and the nonergodic sequences for both the Euclidean and the entropy setup for the simplex constrained least squares problem (51) using $k = 100$, $l = 1000$. While the ergodic sequences both show a $O(1/N^2)$ rate, the nonergodic sequences show a completely different behavior. In the entropy setting, the nonergodic sequence converges a little bit faster but is seems to be quite unstable. In the Euclidean setting, the nonergodic sequence converges extremely fast. We do not know the reason for this, but it will be interesting to find an alternative proof for the convergence rate that does not rely on the ergodic sequence.

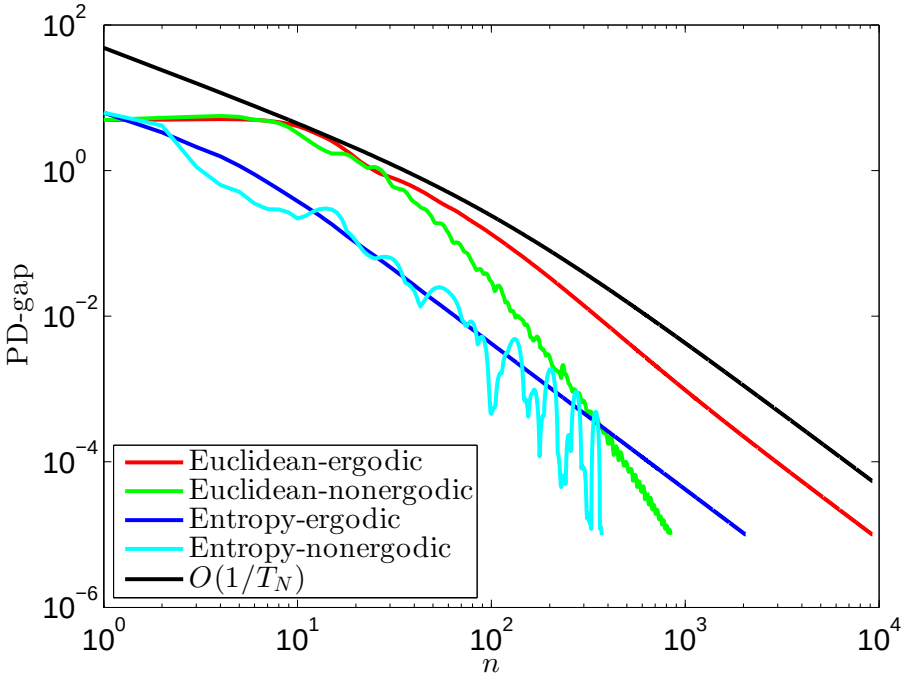

Figure 2: Comparison between the performance of the ergodic and the nonergodic sequences of Algorithm 4 applied to the simplex constrained least squares problem (51).

7.3 Elastic net problem

Finally, we consider the elastic net problem which has been extensively used for feature selection and sparse coding. It is written as the following optimization problem:

$$\min_x \mathcal{P}(x) = \frac{1}{2}\|Kx - b\|^2 + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|^2, \tag{53}$$

where $K \in \mathbb{R}^{k\times l}$ is a matrix where its columns are features and $b \in \mathbb{R}^k$ is the measurement vector. For $\lambda_2 = 0$, the elastic net is equivalent to the well-known LASSO problem. It can be rewritten as the following saddle-point problem:

$$\min_x \max_y \mathcal{L}(x,y) = \langle Kx, y\rangle + \lambda_1\|x\|_1 + \lambda_2\|x\|^2 - \frac{1}{2}\|y\|^2 - b^T y$$

Observe that the above problem is λ_2-strongly convex in x and 1-strongly convex in y. Hence, we can make use of the linearly converging Algorithm 5. The dual problem is computed as

$$\max_y \mathcal{D}(y) = -\frac{1}{2\lambda_2}\|(|K^T y| - \lambda_1)^+\|^2 - \frac{1}{2}\|y\|^2 - b^T y,$$

where the expressions $|K^T y|$ and $(t)^+ = \max(0,t)$ are understood element-wise. In turn the primal-dual gap can be computed as

$$\mathcal{G}(x,y) = \left[\frac{1}{2}\|Kx - b\|^2 + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|^2\right] - \left[-\frac{1}{2\lambda_2}\|(|K^T y| - \lambda_1)^+\|^2 - \frac{1}{2}\|y\|^2 - b^T y\right]. \quad (54)$$

In our experiments, we again choose the entries of K and b uniformly and independently in the interval $[-1,1]$ and we again denote by $L_2 = \|K\| = s_{\max}(K)$ the largest singular value of K. We compute the values for τ, σ and θ using the formulas provided in (49) and we choose $x^0 = 0$, $y^0 = Kx^0 - b$. According to (47), after N iterations, we have for all (x,y):

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) \leq \frac{\|x - x^0\|^2}{2\tau} + \frac{\|y - y^0\|^2}{2\sigma}.$$

Taking the supremum on the left hand with respect to (x,y) we find $x = (|K^T Y^N| - \lambda_1)^+ \cdot \operatorname{sgn}(-K^T Y^N)/\lambda_2$ and $y = KX^N - b$. Substituting back we obtain the final estimate

$$T_N \mathcal{G}(X^N, Y^N) \leq \frac{\|(|K^T Y^N| - \lambda_1)^+\|^2}{2\tau\lambda_2^2} + \frac{L_2^2\|X^N\|^2}{2\sigma} < \infty,$$

where $T_N \sim O(\theta^{-N})$ is defined in (46) and τ, σ are chosen according to (49).

For the implementation of the algorithm we need to solve the proximal map with respect to the mixed ℓ_1-ℓ_2 norm appearing in the primal problem. The solution is given by:

$$\hat{x} = \arg\min_x \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|^2 + \frac{1}{2\tau}\|x - \bar{x}\|^2 \Leftrightarrow \hat{x} = \frac{\max(0, |\bar{x}| - \tau\lambda_1)\cdot \operatorname{sgn}(\bar{x})}{1 + \tau\lambda_2},$$

where the operations are understood element-wise.

In Table 5 we evaluate Algorithm 5 for different problem instances of (53). We set $\lambda_1 = 1$ and used different values of λ_2 in order to study the behavior of the algorithm for different degrees of convexity. The table reports the number of iterations that were needed to achieve a primal-dual gap less than the error tolerance ε. One can see that in general, a smaller value of λ_2 leads to a smaller strong convexity parameter of the primal problem and hence the problem appears more difficult to the algorithm. Thanks to the $O(\theta^N)$ linear convergence rate of the algorithm, reducing the required tolerance by a factor of 10 only leads to a small increase of the required iterations.

Table 5: Computational results of Algorithm 5 applied to the elastic net problem (53).

	$\varepsilon = 10^{-3}$		$\varepsilon = 10^{-4}$	
k/l	$\lambda_2 = 10^{-2}$	$\lambda_2 = 10^{-3}$	$\lambda_2 = 10^{-2}$	$\lambda_2 = 10^{-3}$
100/100	445	1405	577	1823
100/500	446	1339	624	1940
100/1000	459	1319	703	2143
500/100	1015	3209	1227	3879
500/500	1189	3759	1486	4697
500/1000	924	2869	1258	3950
1000/100	1421	4494	1696	5363
1000/500	1753	5542	2109	6667
1000/1000	1707	5397	2123	6714

7.3.1 Ergodic versus nonergodic sequence

Finally Figure 3 shows the performance difference between the ergodic sequence and the nonergodic sequence for the elastic net problem using $k = 100$, $l = 1000$, $\lambda_1 = 1$, and $\lambda_2 = 10^{-3}$. One can see that while the performance of the ergodic sequence is again well predicted by the worst case rate $O(\theta^N)$, the performance of the nonergodic sequence is again superior.

8 Conclusion

In this work, we have presented refined ergodic convergence rates for a first-order primal-dual algorithm for composite convex-concave saddle-point problems. Some of the presented proofs are quite elementary and easily extend to non-linear Bregman distance functions and inertial or overrelaxed variants of the algorithm. Furthermore, we have given refined ergodic convergence rates in terms of the primal-dual gap function for accelerated variants of the algorithm. We have illustrated the theoretical results by applying the algorithms to a number of standard convex optimization problems including matrix games, simplex constrained least squares problems and the elastic net selector. The numerical results show a behaviour which corresponds nicely to the theoretical predictions. We have also observed that in the Euclidean setting, the nonergodic sequences very often converge much faster than the ergodic sequences. We will investigate this issue in more detail in our future research. Furthermore, it will be interesting to investigate strategies to dynamically adjust the step sizes (τ_n, σ_n and θ_n) to the local smoothness of the problem, which can vary a lot during the optimization (see Remark 10).

Acknowledgments

This research is partially supported by the joint ANR/FWF Project *Efficient Algorithms for Nonsmooth Optimization in Imaging (EANOI)* FWF n. I1148 / ANR-12-IS01-0003. A.C. would

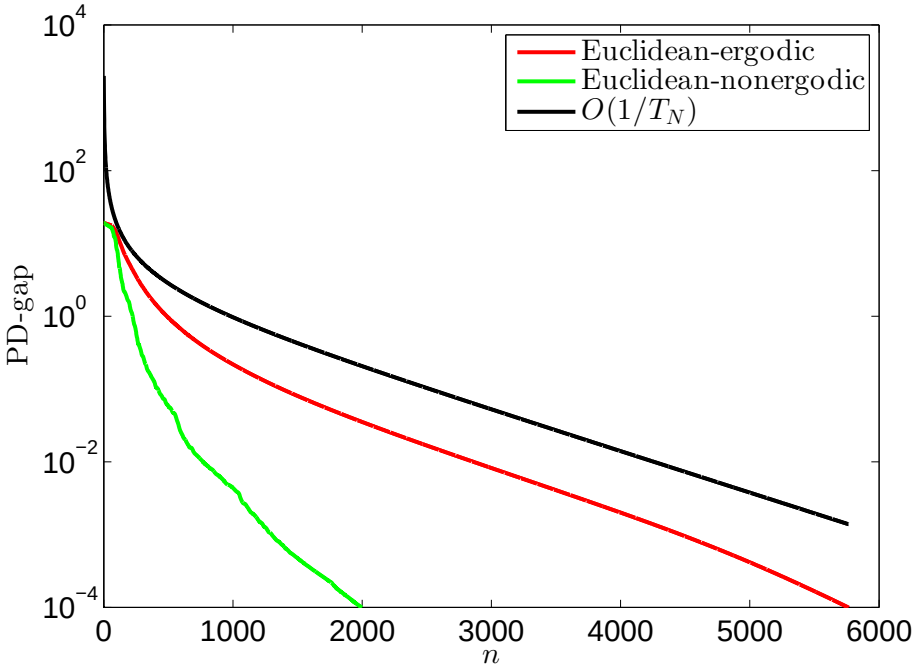

Figure 3: Comparison between the performance of the ergodic and the nonergodic sequences of Algorithm 5 applied to the elastic net problem (53).

like to thank his colleague S. Gaiffas for stimulating discussions, as well as J. Fadili for very helpful discussions on nonlinear proximity operators. This work also benefited from the support of the "Gaspard Monge Program in Optimization and Operations Research" (PGMO), supported by EDF and the Fondation Mathématique Jacques Hadamard (FMJH). The authors also need to thank the referees for their careful reading of the manuscript and their numerous helpful comments.

References

[1] Felipe Alvarez and Hedy Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Anal.*, 9(1-2):3–11, 2001. Wellposedness in optimization and related topics (Gargnano, 1999).

[2] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] Radu Ioan Boţ, Ernö Robert Csetnek, André Heinrich, and Christopher Hendrich. On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. *Math. Program.*, 150(2, Ser. A):251–279, 2015.

[5] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

[6] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3(3):538–543, 1993.

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

[8] Patrick L. Combettes, Laurent Condat, Jean-Christophe Pesquet, and Bằng Công Vũ. A forward-backward view of some primal-dual optimization methods in image recovery. In *Proceedings ICIP 2014 Conference, Paris, Oct. 2014*, 2014. to appear.

[9] Laurent Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, 158(2):460–479, 2013.

[10] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. Technical report, CAM Report 15-13 / preprint arXiv:1504.01032, 2015.

[11] Yoel Drori, Shoham Sabach, and Marc Teboulle. A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. *Oper. Res. Lett.*, 43(2):209–214, 2015.

[12] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 272–279, New York, 2008. ACM.

[13] Jonathan Eckstein. Nonlinear proximal point algorithms using bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[14] Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming*, 55(3, Ser. A):293–318, 1992.

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

[16] Bingsheng He and Xiaoming Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM J. Imaging Sci.*, 5(1):119–149, 2012.

[17] Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.*, 50(2):700–709, 2012.

[18] Thorsten Hohage and Carolin Homann. A generalization of the Chambolle-Pock algorithm to Banach spaces with applications to inverse problems. Technical report, arXiv:1412.0126, 2014.

[19] Dirk A. Lorenz and Thomas Pock. An inertial forward-backward algorithm for monotone inclusions. *J. Math. Imaging Vision*, 51(2):311–325, 2015.

[20] Arkadi S. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251 (electronic), 2004.

[21] Arkadi S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[22] Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[23] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[24] Zdzisław Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bull. Amer. Math. Soc.*, 73:591–597, 1967.

[25] Jean-Christophe Pesquet and Audrey Repetti. A class of randomized primal-dual algorithms for distributed optimization. *arXiv:1406.6404*, 2014.

[26] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *ICCV Proceedings*, LNCS. Springer, 2009.

[27] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14(5):877–898, 1976.

[28] Ron Shefi and Marc Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM J. Optim.*, 24(1):269–297, 2014.

[29] Marc Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Math. Oper. Res.*, 17(3):670–690, 1992.

[30] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008. *Submitted to SIAM J. Optim. / available at* `http://www.csie.ntu.edu.tw/~b97058/tseng/papers/apgm.pdf`.

[31] Bằng Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Comput. Math.*, 38(3):667–681, 2013.