An accelerated minimax algorithm for convex-concave saddle point problems with nonsmooth coupling function

Radu Ioan Boţ[1,2,*], Ernö Robert Csetnek[1,*], and Michael Sedlmayer[2,*]

[1]Faculty of Mathematics
[2]Research Network Data Science @ Uni Vienna
[*]University of Vienna, Austria
{radu.bot, robert.csetnek, michael.sedlmayer}@univie.ac.at

May 13, 2022

Abstract

In this work we aim to solve a convex-concave saddle point problem, where the convex-concave coupling function is smooth in one variable and nonsmooth in the other and *not* assumed to be linear in either. The problem is augmented by a nonsmooth regulariser in the smooth component. We propose and investigate a novel algorithm under the name of *OGAProx*, consisting of an *optimistic gradient ascent* step in the smooth variable coupled with a proximal step of the regulariser, and which is alternated with a *proximal step* in the nonsmooth component of the coupling function. We consider the situations convex-concave, convex-strongly concave and strongly convex-strongly concave related to the saddle point problem under investigation. Regarding iterates we obtain (weak) convergence, a convergence rate of order $\mathcal{O}(\frac{1}{K})$ and linear convergence like $\mathcal{O}(\theta^K)$ with $\theta < 1$, respectively. In terms of function values we obtain ergodic convergence rates of order $\mathcal{O}(\frac{1}{K})$, $\mathcal{O}(\frac{1}{K^2})$ and $\mathcal{O}(\theta^K)$ with $\theta < 1$, respectively. We validate our theoretical considerations on a nonsmooth-linear saddle point problem, the training of multi kernel support vector machines and a classification problem incorporating minimax group fairness.

Key words. saddle point problem, convex-concave, minimax algorithm, convergence rate, acceleration, linear convergence

1 Introduction

Saddle point – or minimax – problems arise traditionally in game theory [23] or for example in the context of determining primal-dual pairs of optimal solutions of constrained convex optimisation problems [1]. However, in recent years they have witnessed increased interest due to many relevant and challenging applications in the field of machine learning, with the most prominent being the training of Generative Adversarial Networks (GANs) [10]. Even though the problems in reality are often not of this form, in the classical setting the minimax objective comprises a smooth convex-concave coupling function with Lipschitz continuous gradient and a (potentially nonsmooth) regulariser in each variable, leading to a convex-concave objective in total.

One well established method in practice due to its simplicity and computational efficiency is *Gradient Descent Ascent* (GDA), either in a *simultaneous* or in an *alternating* variant (for a recent comparison of the convergence behaviour of the two schemes we refer to [24]). However, naive application of GDA is known to lead to oscillatory behaviour or even divergence already in simple cases such as bilinear objectives. Most algorithms with convergence guarantees in the general convex-concave setting make use of the formulation of the first order optimality conditions as monotone inclusion or variational inequality, treating both components in a symmetric fashion. For example we have the *Extragradient method* [12] whose application to minimax problems has been studied in [19] under the name of *Mirror Prox*, and the *Forward-Backward-Forward method* (FBF) [22] with application to saddle point problems in [3]. Both

algorithms have even been successfully applied to the training of GANs (see [9, 3]), but, though being single-loop methods, suffer in practice from requiring two gradient evaluations per iteration. A possible way to avoid this is to reuse previous gradients. Doing this for FBF – as shown in [3] – recovers the *Forward-Reflected Backward method* [15] which was applied to saddle point problems under the name of *Optimistic Mirror Descent* and to GAN training under the name of *Optimistic Gradient Descent Ascent* [6, 5, 14].

The first method treating general coupling functions with an asymmetric scheme is the *Accelerated Primal-Dual Algorithm* (APD) by [11], involving an optimistic gradient ascent step in one component which is followed by a gradient descent step in the other one. In the special case of a bilinear coupling function APD recovers the *Primal-Dual Hybrid Gradient Method* (PDHG) [4]. In the case of the minimax objective being strongly convex-concave acceleration of PDHG is obtained in [4], which is also done for APD in [11], however only under the rather limiting assumption of linearity of the coupling function in one component.

In this paper we introduce a novel algorithm *OGAProx* for solving a convex-concave saddle point problem, where the convex-concave coupling function is smooth in one variable and nonsmooth in the other, and it is augmented by a nonsmooth regulariser in the smooth component. OGAProx consists of an optimistic gradient ascent step in the smooth component of the coupling function combined with a proximal step of the regulariser, which is followed by a proximal step of the coupling function in the nonsmooth component. We will be also able to accelerate our method in the convex-strongly concave setting *without* linearity assumption on the coupling function. Furthermore, we prove linear convergence if the problem is strongly convex-strongly concave, yielding similar results as for PDHG [4] in the bilinear case.

So far in most works nonsmoothness is only introduced via regularisers, as the coupling function is typically accessed through gradient evaluations. Recently there is another development, although with the saddle point problem *not* being convex-concave, where the assumption on differentiability of the coupling function in both components is weakened to only one component [2]. As the evaluation of the proximal mapping does not require differentiability we will assume the coupling function to be smooth in only one component, too.

The remainder of the paper is organised as follows. Next we will introduce the precise problem formulation and the setting we will work with, formulate the proposed algorithm OGAProx and state our contributions. This will be followed by preliminaries in Section 2. Afterwards we will discuss the properties of our algorithm in the convex-concave and convex-strongly concave setting and state respective convergence results in Section 3. After that we will investigate the convergence of the method under the additional assumption of strong convexity-strong concavity in Section 4. The paper will be concluded by numerical experiments in Section 5, where we treat a simple nonsmooth-linear saddle point problem, the training of multi kernel support vector machines and a classification problem taking into account minimax group fairness.

1.1 Problem description

Consider the saddle point problem

$$\min_{x\in\mathcal{H}} \max_{y\in\mathcal{G}} \Psi(x,y) := \Phi(x,y) - g(y), \tag{1}$$

where $\mathcal{H}$, $\mathcal{G}$ are real Hilbert spaces, $\Phi : \mathcal{H}\times\mathcal{G} \to \mathbb{R}\cup\{+\infty\}$ is a coupling function with $\operatorname{dom}\Phi := \{(x,y) \in \mathcal{H}\times\mathcal{G} \mid \Phi(x,y) < +\infty\} \neq \emptyset$ and $g : \mathcal{G} \to \mathbb{R}\cup\{+\infty\}$ a regulariser. Throughout the paper (unless otherwise specified) we will make the following assumptions:

- g is proper, lower semicontinuous and convex with modulus $\nu \geq 0$, i.e. $g - \frac{\nu}{2}\|\cdot\|^2$ is convex (notice that we also allow and consider the situation $\nu = 0$, in which case g is convex; otherwise g is strongly convex);
- for all $y \in \operatorname{dom} g$, $\Phi(\,\cdot\,, y) : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ is proper, convex and lower semicontinuous;

- for all $x \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi) := \{u \in \mathcal{H} \mid \exists y \in \mathcal{G} \text{ such that } (u,y) \in \mathrm{dom}\,\Phi\}$ we have that $\mathrm{dom}\,\Phi(x,\cdot) = \mathcal{G}$ and $\Phi(x,\cdot) : \mathcal{G} \to \mathbb{R}$ is concave and Fréchet differentiable. Moreover, $\mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi)$ is closed;
- there exist L_{yx}, $L_{yy} \geq 0$ such that for all (x,y), $(x',y') \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi) \times \mathrm{dom}\,g$ it holds

$$\left\|\nabla_y \Phi(x,y) - \nabla_y \Phi(x',y')\right\| \leq L_{yx}\left\|x - x'\right\| + L_{yy}\left\|y - y'\right\|. \tag{2}$$

By convention we set $+\infty - (+\infty) := +\infty$. Thus, the situation can be summarised by

$$\Psi(x,y) = \begin{cases} -\infty & \text{if } x \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi) \text{ and } y \notin \mathrm{dom}\,g, \\ \Phi(x,y) - g(y) & \text{if } x \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi) \text{ and } y \in \mathrm{dom}\,g, \\ +\infty & \text{if } x \notin \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi). \end{cases} \tag{3}$$

We are interested in finding a *saddle point* of (1), which is a point $(x^*,y^*) \in \mathcal{H} \times \mathcal{G}$ that fulfils the inequalities

$$\Psi(x^*,y) \leq \Psi(x^*,y^*) \leq \Psi(x,y^*) \quad \forall (x,y) \in \mathcal{H} \times \mathcal{G}. \tag{4}$$

For the remainder we assume that such a saddle point exists.

The assumptions considered above ensure that for any saddle point $(x^*,y^*) \in \mathcal{H} \times \mathcal{G}$ we have

$$x^* \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi), \quad y^* \in \mathrm{dom}\,g \text{ and } \Psi(x^*,y^*) = \Phi(x^*,y^*) - g(y^*) \in \mathbb{R}.$$

Finding a saddle point of (1) amounts to solving the necessary and sufficient first order optimality conditions, given by the following coupled inclusion problems

$$0 \in \partial\left[\Phi(\,\cdot\,,y^*)\right](x^*) \quad \text{and} \quad 0 \in -\nabla_y \Phi(x^*,y^*) + \partial g(y^*).$$

Remark 1. In case Φ and g have full domain, Ψ is a convex-concave function with full domain and the set $\mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi)$ is obviously closed. However, in order to allow more flexibility and to cover a wider range of problems (see also the last section with numerical experiments), our investigations are carried out in the more general setting given by the assumptions described above. Furthermore, these assumptions allow us to stay in the rigorous setting of the theory of convex-concave saddle functions as described by Rockafellar in [21] (see Definition 4 and Proposition 5 below).

Example 2. Consider the nonsmooth convex optimisation problem with inequality constraints

$$\begin{aligned} \min \quad & f(x), \\ \text{subject to} \quad & h_i(x) \leq 0, i = 1,...,m \end{aligned} \tag{5}$$

where $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex and lower semicontinuous function and $h_i : \mathcal{H} \to \mathbb{R}, i = 1,...,m$, are convex and continuous functions. The Lagrangian attached to (5) reads

$$L : \mathcal{H} \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}, \quad L(x,\lambda_1,...,\lambda_m) = f(x) + \sum_{i=1}^m \lambda_i h_i(x).$$

Then the saddle point problem

$$\min_{x \in \mathcal{H}} \max_{(\lambda_1,...,\lambda_m) \in \mathbb{R}^m_+} L(x,\lambda_1,...,\lambda_m) = \min_{x \in \mathcal{H}} \max_{(\lambda_1,...,\lambda_m) \in \mathbb{R}_m} L(x,\lambda_1,...,\lambda_m) - \delta_{\mathbb{R}^+_m}(\lambda_1,...,\lambda_m) \tag{6}$$

exhibits the structure of saddle point problem (1). It is known that if $(x^*,\lambda_1^*,...,\lambda_m^*)$ is a saddle point of (6), then x^* is an optimal solution of the constrained convex optimisation problem (5) and $(\lambda_1^*,...,\lambda_m^*)$ is an optimal solution of its Lagrange dual.

1.2 Algorithm

The algorithm we investigate performs an optimistic gradient ascent step of Φ followed by an evaluation of the proximal mapping g in the variable y, while it carries out a purely proximal step of Φ in x. We will call this method *Optimistic Gradient Ascent – Proximal Point algorithm* (OGAProx) in the following. For all $k \geq 0$ we define

$$\begin{cases} y_{k+1} = \operatorname{prox}_{\sigma_k g}\left(y_k + \sigma_k\left[(1+\theta_k)\nabla_y \Phi(x_k, y_k) - \theta_k \nabla_y \Phi(x_{k-1}, y_{k-1})\right]\right), & (7) \\ x_{k+1} = \operatorname{prox}_{\tau_k \Phi(\,\cdot\,, y_{k+1})}(x_k), & (8) \end{cases}$$

with the conventions $x_{-1} := x_0$ and $y_{-1} := y_0$ for starting points $x_0 \in \operatorname{Pr}_{\mathcal{H}}(\operatorname{dom}\Phi)$ and $y_0 \in \operatorname{dom} g$. The particular choices of the sequences $(\sigma_k)_{k\geq 0}$, $(\tau_k)_{k\geq 0} \subseteq \mathbb{R}_{++}$ and $(\theta_k)_{k\geq 0} \subseteq (0,1]$ will be specified later.

1.3 Contribution

Let us summarize the main results of this paper:

1. We introduce a novel algorithm to solve saddle point problems with nonsmooth coupling functions, which in general is *not* assumed to be linear in either component.
2. We prove for the saddle function Ψ being
   (a) convex-concave (see Theorem 9):
      - weak convergence of the generated sequence $(x_k, y_k)_{k\geq 0}$ to a saddle point (x^*, y^*) as $k \to +\infty$;
      - convergence of the minimax gap $\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)$ to zero like $\mathcal{O}(\frac{1}{K})$ as $K \to +\infty$, where $(\bar{x}_K)_{K\geq 1}$ and $(\bar{y}_K)_{K\geq 1}$ are the *ergodic sequences* obtained by averaging $(x_k)_{k\geq 1}$ and $(y_k)_{k\geq 1}$, respectively;

   (b) convex-strongly concave (see Theorem 12):
      - strong convergence of $(y_k)_{k\geq 0}$ to y^* like $\mathcal{O}(\frac{1}{k})$ as $k \to +\infty$;
      - convergence of the minimax gap $\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)$ to zero like $\mathcal{O}(\frac{1}{K^2})$ as $K \to +\infty$;

   (c) strongly convex-strongly concave (see Theorem 14):
      - linear convergence of $(x_k, y_k)_{k\geq 0}$ to (x^*, y^*) like $\mathcal{O}(\theta^k)$, with $0 < \theta < 1$, as $k \to +\infty$;
      - linear convergence of the minimax gap $\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)$ to zero like $\mathcal{O}(\theta^K)$ as $K \to +\infty$.

2 Preliminaries

We recall some basic notions in convex analysis and monotone operator theory (see for example [1]). The real Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$ are endowed with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{G}}$, respectively. As it will be clear from the context which one is meant, we will drop the index for ease of notation and write $\langle \cdot, \cdot \rangle$ for both. The norm induced by the respective inner products is defined by $\|\cdot\| := \sqrt{\langle \cdot, \cdot \rangle}$.

A function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is said to be *proper* if $\operatorname{dom} f := \{x \in \mathcal{H} : f(x) < +\infty\} \neq \emptyset$. The *(convex) subdifferential* of the function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ at $x \in \mathcal{H}$ is defined by $\partial f(x) := \{u \in \mathcal{H} \mid \langle y - x, u\rangle + f(x) \leq f(y)\ \forall y \in \mathcal{H}\}$ if $f(x) \in \mathbb{R}$ and by $\partial f(x) := \emptyset$ otherwise. If the function f is convex and Fréchet differentiable at $x \in \mathcal{H}$, then $\partial f(x) = \{\nabla f(x)\}$. For the sum of a proper, convex and lower semicontinuous function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and a convex and Fréchet differentiable function $h : \mathcal{H} \to \mathbb{R}$ we have $\partial(f + h)(x) = \partial f(x) + \nabla h(x)$ for all $x \in \mathcal{H}$. The subdifferential of the *indicator function* δ_C of a nonempty closed convex set $C \subseteq \mathcal{H}$, that is defined as $\delta_C(x) = 0$ for $x \in C$ and $\delta_C(x) = +\infty$ otherwise, is denoted by $N_C := \partial \delta_C$ and is called the *normal cone* of the set C.

Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper, convex and lower semicontinuous. The *proximal operator* of f is defined by

$$\text{prox}_f : \mathcal{H} \to \mathcal{H}, \quad \text{prox}_f(x) := \underset{y \in \mathcal{H}}{\arg\min} \left\{ f(y) + \frac{1}{2} \|y - x\|^2 \right\}.$$

The proximal operator of the indicator function δ_C of a nonempty closed convex set $C \subseteq \mathcal{H}$ is the *orthogonal projection* $P_C : \mathcal{H} \to C$ of the set C.

A set-valued operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$ is said to be *monotone* if for all (x, u), $(y, v) \in \text{gra}\, A := \{(z, w) \in \mathcal{H} \times \mathcal{H} \mid w \in Az\}$ we have $\langle x - y, u - v \rangle \geq 0$. Furthermore, A is said to be *maximal monotone* if it is monotone and there exists no monotone operator $B : \mathcal{H} \rightrightarrows \mathcal{H}$ such that $\text{gra}\, A \subsetneqq \text{gra}\, B$. The graph of a maximal monotone operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$ is *sequentially closed* in the *strong* $\times$ *weak* topology, which means that if $(x_k, u_k)_{k \geq 0}$ is a sequence in $\text{gra}\, A$ such that $x_k \to x$ and $u_k \rightharpoonup u$ as $k \to +\infty$, then $(x, u) \in \text{gra}\, A$. The notation $u_k \rightharpoonup u$ as $k \to +\infty$ is used to denote convergence of the sequence $(u_k)_{k \geq 0}$ to u in the weak topology.

To show weak convergence of sequences in Hilbert spaces we use the following so-called *Opial Lemma*.

Lemma 3. *(Opial Lemma [20]) Let $C \subseteq \mathcal{H}$ be a nonempty set and $(x_k)_{k \geq 0}$ a sequence in $\mathcal{H}$ such that the following two conditions hold:*

(a) for every $x \in C$, $\lim_{k \to +\infty} \|x_k - x\|$ exists;

(b) every weak sequential cluster point of $(x_k)_{k \geq 0}$ belongs to C.

Then $(x_k)_{k \geq 0}$ converges weakly to an element in C.

In the following definition we adjust the term *proper* to the saddle point setting and refer to [21] for further considerations related to saddle functions.

Definition 4. *A function $\Psi : \mathcal{H} \times \mathcal{G} \to \mathbb{R} \cup \{\pm\infty\}$ is called a* saddle function*, if $\Psi(\cdot\,, y)$ is convex for all $y \in \mathcal{G}$ and $\Psi(x, \cdot\,)$ is concave for all $x \in \mathcal{G}$. A saddle function Ψ is called* proper*, if there exists $(x', y') \in \mathcal{H} \times \mathcal{G}$ such that $\Psi(x', y) < +\infty$ for all $y \in \mathcal{G}$ and $-\infty < \Psi(x, y')$ for all $x \in \mathcal{H}$.*

We conclude the preliminary section with a useful result regarding the minimax objective from (1).

Proposition 5. *The function $\Psi : \mathcal{H} \times \mathcal{G} \to \mathbb{R} \cup \{\pm\infty\}$ defined via (3) is a proper saddle function such that $\Psi(\cdot\,, y)$ is lower semicontinuous for each $y \in \mathcal{G}$ and $\Psi(x, \cdot\,)$ is upper semicontinuous for each $x \in \mathcal{H}$. Consequently, the operator*

$$(x, y) \mapsto \partial[\Psi(\cdot\,, y)](x) \times \partial[-\Psi(x, \cdot\,)](y)$$

is maximal monotone.

Proof. We choose $(x', y') \in \mathcal{H} \times \mathcal{G}$ and distinguish four cases.

Firstly, we look at the case $y' \notin \text{dom}\, g$. Then

$$\Psi(x, y') = \begin{cases} -\infty & \text{if } x \in \text{Pr}_{\mathcal{H}}(\text{dom}\, \Phi), \\ +\infty & \text{if } x \notin \text{Pr}_{\mathcal{H}}(\text{dom}\, \Phi), \end{cases}$$

thus $x \mapsto \Psi(x, y')$ is convex and lower semicontinuous, since $\text{Pr}_{\mathcal{H}}(\text{dom}\, \Phi)$ is convex and closed. Secondly, if $y' \in \text{dom}\, g$, then $g(y') \in \mathbb{R}$ and

$$\Psi(x, y') = \Phi(x, y') - g(y') \; \forall x \in \mathcal{H},$$

which means that $x \mapsto \Psi(x, y')$ is convex and lower semicontinuous. This proves that $\Psi(\cdot\,, y)$ is convex and lower semicontinuous for all $y \in \mathcal{G}$.

On the other hand, if $x' \notin \text{Pr}_{\mathcal{H}}(\text{dom}\, \Phi)$, then

$$\Psi(x', y) = +\infty \; \forall y \in \mathcal{G},$$

which means that $y \mapsto \Psi(x', y)$ is upper semicontinuous and concave. Finally, if $x' \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi)$, then

$$\Phi(x', y) \in \mathbb{R} \text{ and } -\Psi(x', y) = -\Phi(x', y) + g(y) \ \forall y \in \mathcal{G}.$$

Hence $y \mapsto -\Psi(x', y)$ is proper, convex and lower semicontinuous, and so $y \mapsto \Psi(x', y)$ is concave and upper semicontinuous. This proves that $\Psi(x, \cdot\,)$ is concave and upper semicontinuous for all $x \in \mathcal{H}$.

Moreover, Ψ is a proper saddle function. By assumption we have $g(y) > -\infty$ for all $y \in \mathcal{G}$ and there exists $x' \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi) \neq \emptyset$ such that $\mathrm{dom}\,\Phi(x', \cdot\,) = \mathcal{G}$. Thus

$$\Psi(x', y) = \Phi(x', y) - g(y) < +\infty \ \forall y \in \mathcal{G}.$$

Furthermore, by assumption there exist $y' \in \mathrm{dom}\, g \subseteq \mathcal{G}$ such that $g(y') < +\infty$ and for all $x \in \mathcal{H}$ we have $\Phi(x, y') > -\infty$. Hence,

$$\Psi(x, y') = \Phi(x, y') - g(y') > -\infty \ \forall x \in \mathcal{H}.$$

The maximal monotonicity of $(x, y) \mapsto \partial[\Psi(\,\cdot\,, y)](x) \times \partial[-\Psi(x, \cdot\,)](y)$ follows from Corollary 1 and Theorem 3 in [21, pages 248-249]. □

3 Convex-(strongly) concave setting

First we will treat the case when the coupling function Φ is convex-concave and g is convex with modulus $\nu \geq 0$. In the case $\nu = 0$ this corresponds to $\Psi(x, y) = \Phi(x, y) - g(y)$ being convex-concave, while for $\nu > 0$ the saddle function $\Psi(x, y)$ is convex-strongly concave.

We will start with stating two assumptions on the step sizes of the algorithm which will be needed in the convergence analysis. These will be followed by a unified preparatory analysis for general $\nu \geq 0$ that will be the base to show convergence of the iterates as well as of the minimax gap. After that we will introduce a choice of parameters that satisfy the aforementioned assumptions. The section will be closed by convergence results for the convex-concave ($\nu = 0$) and the convex-strongly concave ($\nu > 0$) setting.

Assumption 1. *We assume that the step sizes τ_k, σ_k and the momentum parameter θ_k satisfy*

$$\tau_{k+1} \geq \frac{\tau_k}{\theta_{k+1}} \quad \text{and} \quad \sigma_{k+1} \geq \frac{\sigma_k}{\theta_{k+1}(1 + \nu\sigma_k)} \quad \text{for all } k \geq 0. \tag{9}$$

Furthermore, we assume that there exist $\delta > 0$ and $(\alpha_k)_{k \geq 0} \subseteq \mathbb{R}_{++}$ such that

$$\frac{1 - \delta}{\tau_k} \geq \frac{L_{yx}}{\alpha_{k+1}} \quad \text{and} \quad \frac{1 - \delta}{\sigma_k} \geq L_{yx}\alpha_k\theta_k + L_{yy}(1 + \theta_k) \quad \text{for all } k \geq 0, \tag{10}$$

where $\theta_0 := 1$.

3.1 Preliminary considerations

In this subsection we will make some preliminary considerations that will play an important role when proving the convergence properties of the numerical scheme given by (7)-(8). For all $k \geq 0$ we will use the notations

$$q_k := \nabla_y \Phi(x_k, y_k) - \nabla_y \Phi(x_{k-1}, y_{k-1}) \text{ and } s_k := \theta_k q_k + \nabla_y \Phi(x_k, y_k). \tag{11}$$

We take an arbitrary $(x, y) \in \mathcal{H} \times \mathcal{G}$ and let $k \geq 0$ be fixed. From (7) we derive

$$0 \in \partial g(y_{k+1}) + \frac{1}{\sigma_k}(y_{k+1} - y_k) - s_k, \tag{12}$$

and, as g is convex with modulus ν, this implies

$$\begin{aligned} g(y) &\geq g(y_{k+1}) + \langle s_k, y - y_{k+1} \rangle + \frac{1}{\sigma_k} \langle y_k - y_{k+1}, y - y_{k+1} \rangle + \frac{\nu}{2} \|y - y_{k+1}\|^2 \\ &= g(y_{k+1}) + \langle s_k, y - y_{k+1} \rangle + \frac{1}{2\sigma_k} \left(\|y_k - y_{k+1}\|^2 + \|y - y_{k+1}\|^2 - \|y - y_k\|^2 \right) + \frac{\nu}{2} \|y - y_{k+1}\|^2 . \end{aligned} \tag{13}$$

From (8) we get

$$0 \in \partial[\Phi(\cdot, y_{k+1})](x_{k+1}) + \frac{1}{\tau_k}(x_{k+1} - x_k), \tag{14}$$

hence the convexity of $\Phi(\cdot, y)$ for $y \in \operatorname{dom} g$ yields

$$\begin{aligned} \Phi(x, y_{k+1}) &\geq \Phi(x_{k+1}, y_{k+1}) + \frac{1}{\tau_k} \langle x_k - x_{k+1}, x - x_{k+1} \rangle \\ &= \Phi(x_{k+1}, y_{k+1}) + \frac{1}{2\tau_k} \left(\|x_k - x_{k+1}\|^2 + \|x - x_{k+1}\|^2 - \|x - x_k\|^2 \right). \end{aligned} \tag{15}$$

Combining (13) and (15) we obtain

$$\begin{aligned} \Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) &= \Phi(x_{k+1}, y) - g(y) - \Phi(x, y_{k+1}) + g(y_{k+1}) \\ &\leq \Phi(x_{k+1}, y) - \Phi(x, y_{k+1}) + \langle s_k, y_{k+1} - y \rangle - \frac{\nu}{2} \|y - y_{k+1}\|^2 \\ &\quad + \frac{1}{2\sigma_k} \left(- \|y_k - y_{k+1}\|^2 - \|y - y_{k+1}\|^2 + \|y - y_k\|^2 \right) \\ &\leq \Phi(x_{k+1}, y) - \Phi(x_{k+1}, y_{k+1}) + \langle s_k, y_{k+1} - y \rangle - \frac{\nu}{2} \|y - y_{k+1}\|^2 \\ &\quad + \frac{1}{2\tau_k} \left(- \|x_k - x_{k+1}\|^2 - \|x - x_{k+1}\|^2 + \|x - x_k\|^2 \right) \\ &\quad + \frac{1}{2\sigma_k} \left(- \|y_k - y_{k+1}\|^2 - \|y - y_{k+1}\|^2 + \|y - y_k\|^2 \right), \end{aligned}$$

which, together with the concavity of Φ in the second variable and (11), gives

$$\begin{aligned} \Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) &\leq \theta_k \langle q_k, y_{k+1} - y \rangle - \frac{\nu}{2} \|y - y_{k+1}\|^2 \\ &\quad - \langle \nabla_y \Phi(x_{k+1}, y_{k+1}), y_{k+1} - y \rangle + \langle \nabla_y \Phi(x_k, y_k), y_{k+1} - y \rangle \\ &\quad + \frac{1}{2\tau_k} \left(- \|x_k - x_{k+1}\|^2 - \|x - x_{k+1}\|^2 + \|x - x_k\|^2 \right) \\ &\quad + \frac{1}{2\sigma_k} \left(- \|y_k - y_{k+1}\|^2 - \|y - y_{k+1}\|^2 + \|y - y_k\|^2 \right) \\ &= - \langle q_{k+1}, y_{k+1} - y \rangle + \theta_k \langle q_k, y_k - y \rangle - \frac{\nu}{2} \|y - y_{k+1}\|^2 \\ &\quad + \frac{1}{2\tau_k} \left(- \|x_k - x_{k+1}\|^2 - \|x - x_{k+1}\|^2 + \|x - x_k\|^2 \right) \\ &\quad + \frac{1}{2\sigma_k} \left(- \|y_k - y_{k+1}\|^2 - \|y - y_{k+1}\|^2 + \|y - y_k\|^2 \right) \\ &\quad + \theta_k \langle q_k, y_{k+1} - y_k \rangle. \end{aligned} \tag{16}$$

By using (2) we can evaluate the last term in the above expression as follows

$$\begin{aligned} |\langle q_k, y - y_k \rangle| &\leq \|q_k\| \, \|y - y_k\| \leq (L_{yx} \|x_k - x_{k-1}\| + L_{yy} \|y_k - y_{k-1}\|) \|y - y_k\| \\ &\leq \frac{L_{yx}}{2} \left(\alpha_k \|y - y_k\|^2 + \frac{1}{\alpha_k} \|x_k - x_{k-1}\|^2 \right) + \frac{L_{yy}}{2} \left(\|y - y_k\|^2 + \|y_k - y_{k-1}\|^2 \right), \end{aligned} \tag{17}$$

with $\alpha_k > 0$ chosen such that (10) holds.

Writing (17) for $y := y_{k+1}$ and combining the resulting inequality with (16) we derive

$$\Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) \leq a_k(x, y) - b_{k+1}(x, y) - c_k, \tag{18}$$

where

$$\begin{aligned} a_k(x, y) := \frac{1}{2\tau_k} \|x - x_k\|^2 + \frac{1}{2\sigma_k} \|y - y_k\|^2 + \theta_k \langle q_k, y_k - y \rangle + \theta_k \frac{L_{yx}}{2\alpha_k} \|x_k - x_{k-1}\|^2 \\ + \theta_k \frac{L_{yy}}{2} \|y_k - y_{k-1}\|^2, \end{aligned}$$

$$b_{k+1}(x,y) := \frac{1}{2\tau_k}\|x - x_{k+1}\|^2 + \frac{1}{2}\left(\frac{1}{\sigma_k} + \nu\right)\|y - y_{k+1}\|^2 + \langle q_{k+1}, y_{k+1} - y\rangle$$
$$+ \frac{L_{yx}}{2\alpha_{k+1}}\|x_{k+1} - x_k\|^2 + \frac{L_{yy}}{2}\|y_{k+1} - y_k\|^2,$$

and

$$c_k := \frac{1}{2}\left(\frac{1}{\tau_k} - \frac{L_{yx}}{\alpha_{k+1}}\right)\|x_{k+1} - x_k\|^2 + \frac{1}{2}\left(\frac{1}{\sigma_k} - L_{yy} - \theta_k(L_{yx}\alpha_k + L_{yy})\right)\|y_{k+1} - y_k\|^2.$$

Now, let us define for all $k \geq 0$

$$t_k := \frac{\theta_0}{\theta_0\theta_1\cdots\theta_k} \tag{19}$$

and notice that

$$\frac{t_k}{t_{k+1}} = \theta_{k+1}.$$

Relation (9) from Assumption 1 is equivalent to

$$\frac{t_k}{\tau_k} \geq \frac{t_{k+1}}{\tau_{k+1}} \quad \text{and} \quad t_k\left(\frac{1}{\sigma_k} + \nu\right) \geq \frac{t_{k+1}}{\sigma_{k+1}}, \tag{20}$$

which will be used in telescoping arguments in the following.

Let $K \geq 1$ and denote

$$T_K := \sum_{k=0}^{K-1} t_k, \quad \bar{x}_K := \frac{1}{T_K}\sum_{k=0}^{K-1} t_k x_{k+1}, \quad \bar{y}_K := \frac{1}{T_K}\sum_{k=0}^{K-1} t_k y_{k+1}. \tag{21}$$

Multiplying both sides of (18) by $t_k > 0$ as defined in (19), followed by summing up the inequalities for $k = 0, \ldots, K-1$ gives

$$\sum_{k=0}^{K-1} t_k\left(\Psi(x_{k+1}, y) - \Psi(x, y_{k+1})\right) \leq \sum_{k=0}^{K-1} t_k\left(a_k(x,y) - b_{k+1}(x,y) - c_k\right).$$

By Jensen's inequality, as $\Psi(\,\cdot\,, y) - \Psi(x, \,\cdot\,)$ is a convex function, we obtain

$$T_K\left(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)\right) \leq \sum_{k=0}^{K-1} t_k\left(\Psi(x_{k+1}, y) - \Psi(x, y_{k+1})\right),$$

and thus

$$T_K\left(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)\right) \leq \sum_{k=0}^{K-1} t_k\left(a_k(x,y) - b_{k+1}(x,y) - c_k\right). \tag{22}$$

Furthermore, using (20), we get for all $k \geq 0$

$$\begin{aligned}
t_k b_{k+1}(x,y) &= \frac{t_k}{2\tau_k}\|x - x_{k+1}\|^2 + \frac{t_k}{2}\left(\frac{1}{\sigma_k} + \nu\right)\|y - y_{k+1}\|^2 + t_k\langle q_{k+1}, y_{k+1} - y\rangle \\
&\quad + t_k\frac{L_{yx}}{2\alpha_{k+1}}\|x_{k+1} - x_k\|^2 + t_k\frac{L_{yy}}{2}\|y_{k+1} - y_k\|^2 \\
&\geq \frac{t_{k+1}}{2\tau_{k+1}}\|x - x_{k+1}\|^2 + \frac{t_{k+1}}{2\sigma_{k+1}}\|y - y_{k+1}\|^2 + t_{k+1}\theta_{k+1}\langle q_{k+1}, y_{k+1} - y\rangle \\
&\quad + t_{k+1}\theta_{k+1}\frac{L_{yx}}{2\alpha_{k+1}}\|x_{k+1} - x_k\|^2 + t_{k+1}\theta_{k+1}\frac{L_{yy}}{2}\|y_{k+1} - y_k\|^2 \\
&= t_{k+1}a_{k+1}(x,y).
\end{aligned}$$

Notice that by (10) in Assumption 1 there exists $\delta > 0$ such that for all $k \geq 0$

$$c_k \geq \delta \left(\frac{1}{2\tau_k} \|x_{k+1} - x_k\|^2 + \frac{1}{2\sigma_k} \|y_{k+1} - y_k\|^2 \right) \geq 0. \tag{23}$$

For the following recall that $x_{-1} = x_0$ and $y_{-1} = y_0$, which implies $q_0 = 0$. By using the above two inequalities in (22) and writing (17) for $k = K$ we obtain

$$\begin{aligned}
\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K) &\leq \frac{1}{T_K} \sum_{k=0}^{K-1} (t_k a_k(x,y) - t_{k+1} a_{k+1}(x,y)) = \frac{1}{T_K} (t_0 a_0(x,y) - t_K a_K(x,y)) \\
&= \frac{1}{T_K} \left(\frac{t_0}{2\tau_0} \|x - x_0\|^2 + \frac{t_0}{2\sigma_0} \|y - y_0\|^2 \right) - \frac{t_K}{T_K} \left(\frac{1}{2\tau_K} \|x - x_K\|^2 + \frac{1}{2\sigma_K} \|y - y_K\|^2 \right) \\
&\quad - \frac{t_K \theta_K}{T_K} \left(\langle q_K, y_K - y \rangle + \frac{L_{yx}}{2\alpha_K} \|x_K - x_{K-1}\|^2 + \frac{L_{yy}}{2} \|y_K - y_{K-1}\|^2 \right) \\
&\leq \frac{1}{T_K} \left(\frac{t_0}{2\tau_0} \|x - x_0\|^2 + \frac{t_0}{2\sigma_0} \|y - y_0\|^2 \right) \\
&\quad - \frac{t_K}{T_K} \left(\frac{1}{2\tau_K} \|x - x_K\|^2 + \frac{1}{2} \left(\frac{1}{\sigma_K} - \theta_K (L_{yx} \alpha_K + L_{yy}) \right) \|y - y_K\|^2 \right).
\end{aligned} \tag{24}$$

By definition we have $t_0 = 1$ and by (10) that the last term of the above inequality is nonpositive, hence the following estimate for the minimax gap function evaluated at the ergodic sequences holds

$$\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K) \leq \frac{1}{T_K} \left(\frac{1}{2\tau_0} \|x - x_0\|^2 + \frac{1}{2\sigma_0} \|y - y_0\|^2 \right) \quad \forall K \geq 1. \tag{25}$$

With these considerations at hand – in specific we want to point out (18), (24) and (25) – we will be able to obtain convergence statements for the two settings $\nu = 0$ and $\nu > 0$.

3.2 Fulfilment of step size assumptions

In this subsection we will investigate a particular choice of parameters to fulfil Assumption 1 which is suitable for both cases of $\nu = 0$ and $\nu > 0$.

Proposition 6. *Let $\nu \geq 0$, $c_\alpha > L_{yx} \geq 0$, $\theta_0 = 1$ and $\tau_0, \sigma_0 > 0$ such that*

$$(c_\alpha L_{yx} \tau_0 + 2L_{yy}) \sigma_0 < 1.$$

We define

$$\theta_{k+1} := \frac{1}{\sqrt{1 + \nu\sigma_k}}, \quad \tau_{k+1} := \frac{\tau_k}{\theta_{k+1}}, \quad \sigma_{k+1} := \theta_{k+1}\sigma_k \quad \text{for all } k \geq 0. \tag{26}$$

Then the sequence $(\tau_k)_{k\geq 0}$, $(\sigma_k)_{k\geq 0}$ and $(\theta_k)_{k\geq 0}$ fulfil (9) in Assumption 1 with equality and (10) for

$$\alpha_k := \begin{cases} c_\alpha \tau_0 & \text{if } k = 0, \\ c_\alpha \tau_{k-1} & \text{if } k \geq 1, \end{cases} \tag{27}$$

and

$$\delta := \min \left\{ 1 - \frac{L_{yx}}{c_\alpha}, 1 - (c_\alpha L_{yx} \tau_0 + 2L_{yy}) \sigma_0 \right\} > 0. \tag{28}$$

Furthermore, for $(t_k)_{k\geq 0}$ defined as in (19) we have

$$t_k = \frac{\theta_0}{\theta_0 \theta_1 \cdots \theta_k} = \frac{\tau_k}{\tau_0} \quad \forall k \geq 0. \tag{29}$$

Proof. First, we show that the particular choice (26) fulfils (9) in Assumption 1 with equality. We see that for all $k \geq 0$

$$\tau_{k+1} = \frac{\tau_k}{\theta_{k+1}},$$

as well as

$$\sigma_{k+1} = \theta_{k+1}\sigma_k = \frac{\sigma_k}{\theta_{k+1}\frac{1}{\theta_{k+1}^2}} = \frac{\sigma_k}{\theta_{k+1}(1+\nu\sigma_k)},$$

follow straight forward by definition.

Next, we show that (10) in Assumption 1 holds for δ defined in (28) with the choices (26) and (27). The first inequality of (10) is equivalent to

$$1 - \delta \geq \frac{L_{yx}}{\alpha_{k+1}}\tau_k = \frac{L_{yx}}{c_\alpha} \quad \forall k \geq 0,$$

which clearly is fulfilled as

$$\delta \leq 1 - \frac{L_{yx}}{c_\alpha}.$$

On the other hand, the second inequality of (10) is equivalent to

$$1 - \delta \geq L_{yx}\alpha_k\theta_k\sigma_k + L_{yy}(1+\theta_k)\sigma_k \quad \forall k \geq 0.$$

By definition of the step size parameters (26) we have for all $k \geq 0$

$$\tau_{k+1}\sigma_{k+1} = \tau_0\sigma_0, \quad \theta_{k+1} \leq 1 = \theta_0, \quad \sigma_{k+1} \leq \sigma_0, \quad \theta_{k+1}\tau_{k+1} = \tau_k,$$

and thus

$$\begin{aligned} 1 - \delta &\geq L_{yx}\alpha_0\theta_0\sigma_0 + L_{yy}(1+\theta_0)\sigma_0 = c_\alpha L_{yx}\tau_0\sigma_0 + 2L_{yy}\sigma_0 \geq c_\alpha L_{yx}\theta_{k+1}^2\tau_{k+1}\sigma_{k+1} + L_{yy}(1+\theta_{k+1})\sigma_{k+1} \\ &= L_{yx}c_\alpha\tau_k\theta_{k+1}\sigma_{k+1} + L_{yy}(1+\theta_{k+1})\sigma_{k+1} = L_{yx}\alpha_{k+1}\theta_{k+1}\sigma_{k+1} + L_{yy}(1+\theta_{k+1})\sigma_{k+1}. \end{aligned}$$

This chain of inequalities holds since

$$\delta \leq 1 - \left(c_\alpha L_{yx}\tau_0 + 2L_{yy}\right)\sigma_0.$$

Finally, using the definition of t_k and (26) we conclude that for all $k \geq 0$

$$t_k = \frac{\theta_0}{\theta_0\theta_1\cdots\theta_k} = \frac{\frac{\tau_0}{\tau_0}}{\frac{\tau_0}{\tau_0}\frac{\tau_0}{\tau_1}\cdots\frac{\tau_{k-1}}{\tau_k}} = \frac{\tau_k}{\tau_0}.$$

□

Remark 7. The choice $L_{yy} = 0$ in (2) which was considered in [11] in the convex-strongly concave setting corresponds to the case when the coupling function Φ is linear in y. We will prove convergence also for L_{yy} positive, which makes our algorithm applicable to a much wider range of problems, as we will see in the section with the numerical experiments.

When the coupling function $\Phi : \mathcal{H} \times \mathcal{G} \to \mathbb{R}$ is bilinear, that is $\Phi(x,y) = \langle y, Ax\rangle$ for some nonzero continuous linear operator $A : \mathcal{H} \to \mathcal{G}$ then we are in the setting of [4]. In this situation one can choose $L_{yy} = 0$ and $L_{yx} = \|A\|$, and (28) yields

$$\delta = \min\left\{1 - \frac{\|A\|}{c_\alpha},\, 1 - c_\alpha\|A\|\tau_0\sigma_0\right\},$$

with $c_\alpha > \|A\|$. To guarantee $\delta > 0$ we fix $0 < \varepsilon < 1$ and set

$$c_\alpha = (1-\varepsilon)^{-1}\|A\|.$$

Hence, we need to satisfy

$$\tau_0\sigma_0\|A\|^2 < 1 - \varepsilon,$$

which heavily resembles the step size condition of [4, Algorithm 2]. Since $\operatorname{prox}_{\gamma\Phi(\cdot,y)}(x) = x - \gamma A^* y$ for all $(x,y) \in \mathcal{H} \times \mathcal{G}$ and all $\gamma > 0$, our OGAProx scheme becomes the primal-dual algorithm PDHG from [4].

3.3 Convergence results

In this subsection we combine the preliminary considerations with the choice of parameters (26) from Proposition 6.

We will start with the case $\nu = 0$ and constant step sizes, which gives weak convergence of the iterates to a saddle point (x^*, y^*) and convergence of the minimax gap evaluated at the ergodic iterates to zero like $\mathcal{O}(\frac{1}{K})$. Afterwards we will consider the case $\nu > 0$, which leads to an accelerated version of the algorithm with improved convergence results. In this setting we obtain convergence of $(y_k)_{k\geq 0}$ to y^* like $\mathcal{O}(\frac{1}{K})$ and convergence of the minimax gap evaluated at the ergodic iterates to zero like $\mathcal{O}(\frac{1}{K^2})$.

3.3.1 Convex-concave setting

For the following we assume that the function g is convex with modulus $\nu = 0$, meaning it is merely convex. Using the results of the previous subsection we will show that with the choice (26) all the parameters are constant.

Proposition 8. *Let $c_\alpha > L_{yx} \geq 0$ and $\tau, \sigma > 0$ such that*

$$(c_\alpha L_{yx}\tau + 2L_{yy})\,\sigma < 1.$$

If $\nu = 0$, then the sequences $(\tau_k)_{k\geq 0}$, $(\sigma_k)_{k\geq 0}$ and $(\theta_k)_{k\geq 0}$ as defined in Proposition 6 are constant, in particular we have

$$\tau_k = \tau_0 := \tau, \quad \sigma_k = \sigma_0 := \sigma, \quad \theta_k = \theta_0 = 1 \quad \text{for all } k \geq 0. \tag{30}$$

Proof. As $\nu = 0$, (26) gives for all $k \geq 0$

$$\theta_{k+1} = \frac{1}{\sqrt{1+\nu\sigma_k}} = 1, \quad \tau_{k+1} = \frac{\tau_k}{\theta_{k+1}} = \tau_0, \quad \sigma_{k+1} = \theta_{k+1}\sigma_k = \sigma_0.$$

□

Next we will state and prove the convergence results in the convex-concave case.

Theorem 9. *Let $c_\alpha > L_{yx} \geq 0$ and $\tau, \sigma > 0$ such that*

$$(c_\alpha L_{yx}\tau + 2L_{yy})\,\sigma < 1.$$

Then the sequence $(x_k, y_k)_{k\geq 0}$ generated by OGAProx with the choice of constant parameters as in Proposition 8, namely,

$$\tau_k = \tau_0 := \tau, \quad \sigma_k = \sigma_0 := \sigma, \quad \theta_k = \theta_0 = 1 \quad \text{for all } k \geq 0,$$

converges weakly to a saddle point $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ of (1). Furthermore, let $K \geq 1$ and denote

$$\bar{x}_K = \frac{1}{K}\sum_{k=0}^{K-1} x_{k+1} \quad \text{and} \quad \bar{y}_K = \frac{1}{K}\sum_{k=0}^{K-1} y_{k+1}.$$

Then for all $K \geq 1$ and any saddle point $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ of (1) we have

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right).$$

Proof. First we will show weak convergence of the sequence of iterates $(x_k, y_k)_{k\geq 0}$ to some saddle point $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ of (1). For this we will use the Opial Lemma (see Lemma 3).

Let $k \geq 0$ and $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ be an arbitrary but fixed saddle point. From (18) together with the choice (30) of constant parameters $\theta_k = 1$, $\tau_k = \tau$, $\sigma_k = \sigma$ and $\alpha_k = \alpha$ we obtain

$$0 \leq \Psi(x_{k+1}, y^*) - \Psi(x^*, y_{k+1}) \leq a_k(x^*, y^*) - b_{k+1}(x^*, y^*) - c_k = a_k(x^*, y^*) - a_{k+1}(x^*, y^*) - c_k, \tag{31}$$

since

$$\begin{aligned} a_k(x^*, y^*) &= \frac{1}{2\tau}\|x^* - x_k\|^2 + \frac{1}{2\sigma}\|y^* - y_k\|^2 + \langle q_k, y_k - y^*\rangle + \frac{L_{yx}}{2\alpha}\|x_k - x_{k-1}\|^2 + \frac{L_{yy}}{2}\|y_k - y_{k-1}\|^2 \\ &= b_k(x^*, y^*), \end{aligned} \tag{32}$$

and

$$c_k = \frac{1}{2}\left(\frac{1}{\tau} - \frac{L_{yx}}{\alpha}\right)\|x_{k+1} - x_k\|^2 + \frac{1}{2}\left(\frac{1}{\sigma} - L_{yy} - (L_{yx}\alpha + L_{yy})\right)\|y_{k+1} - y_k\|^2.$$

We see that (32), writing (17) with $y = y^*$ and (9) in Assumption 1 yield

$$a_k(x^*, y^*) \geq \frac{1}{2\tau}\|x^* - x_k\|^2 + \frac{1}{2\sigma}\left(1 - \sigma(L_{yx}\alpha + L_{yy})\right)\|y^* - y_k\|^2 \geq 0. \tag{33}$$

Furthermore, from (31) and (23) we deduce

$$a_k(x^*, y^*) \geq a_{k+1}(x^*, y^*) + \delta\left(\frac{1}{2\tau}\|x_{k+1} - x_k\|^2 + \frac{1}{2\sigma}\|y_{k+1} - y_k\|^2\right).$$

Telescoping this inequality and taking into account (33) give

$$\lim_{k\to+\infty}(x_{k+1} - x_k) = \lim_{k\to+\infty}(y_{k+1} - y_k) = 0, \tag{34}$$

as well as the existence of the limit $\lim_{k\to+\infty} a_k(x^*, y^*) \in \mathbb{R}$.

From (33) we get that $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ are bounded sequences. Moreover, by using (2) and (34) in definition (11) we obtain that

$$(q_k)_{k\geq 0} \text{ converges strongly to } 0. \tag{35}$$

From the definition of $a_k(x^*, y^*)$ in (32), (34) and (35) we derive that

$$\exists \lim_{k\to+\infty}\left(\frac{1}{2\tau}\|x_k - x^*\|^2 + \frac{1}{2\sigma}\|y_k - y^*\|^2\right) \in \mathbb{R}.$$

Since this is true for an arbitrary saddle point $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$, we have that the first statement of the Opial Lemma holds.

Next we will show that all weak cluster points of $(x_k, y_k)_{k\geq 0}$ are in fact saddle points of (1). Assume that $(x_{k_n})_{n\geq 0}$ converges weakly to $x^* \in \mathcal{H}$ and $(y_{k_n})_{n\geq 0}$ converges weakly to $y^* \in \mathcal{G}$ as $n \to +\infty$. From (14), (11) and (12) we have

$$\begin{aligned} &\left(\frac{1}{\tau}(x_{k_n} - x_{k_n+1}), \frac{1}{\sigma}(y_{k_n} - y_{k_n+1}) + q_{k_n} - q_{k_n+1}\right) \\ &\qquad \in \partial[\Phi(\,\cdot\,, y_{k_n+1})](x_{k_n+1}) \times (-\nabla_y \Phi(x_{k_n+1}, y_{k_n+1}) + \partial g(y_{k_n+1})) \\ &\qquad = \partial[\Psi(\,\cdot\,, y_{k_n+1})](x_{k_n+1}) \times \partial[-\Psi(x_{k_n+1}, \,\cdot\,)](y_{k_n+1}), \end{aligned} \tag{36}$$

where we used that for all $k \geq 0$ we have $x_k \in \mathrm{Pr}_{\mathcal{H}}(\mathrm{dom}\,\Phi)$ and $y_k \in \mathrm{dom}\, g$. The sequence on the left hand side of the inclusion (36) converges strongly to $(0, 0)$ as $n \to +\infty$ (according to (34) and (35)). Notice that the operator $(x, y) \mapsto \partial[\Psi(\,\cdot\,, y)](x) \times \partial[-\Psi(x, \,\cdot\,)](y)$ is maximal monotone (see Proposition 5), hence its graph is sequentially closed with respect to the strong $\times$ weak topology. From here we deduce

$$(0, 0) \in \partial[\Psi(\,\cdot\,, y^*)](x^*) \times \partial[-\Psi(x^*, \,\cdot\,)](y^*),$$

from which we easily derive that (x^*, y^*) is a saddle point as it satisfies (4). This means that also the second statement of the Opial Lemma is fulfilled and we have weak convergence of $(x_k, y_k)_{k\geq 0}$ to a saddle point (x^*, y^*).

The remaining part is to show the convergence rate of the minimax gap of the ergodic sequences. Let $K \geq 1$ and $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ be an arbitrary but fixed saddle point. Writing (25) for (x^*, y^*) yields

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{T_K}\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right),$$

with

$$T_K = \sum_{k=0}^{K-1} t_k, \quad \bar{x}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k x_{k+1}, \quad \bar{y}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k y_{k+1}.$$

Using (29) to get $t_k = 1$ for all $k \geq 0$ in the above expressions gives

$$T_K = \sum_{k=0}^{K-1} t_k = K, \quad \bar{x}_K = \frac{1}{K}\sum_{k=0}^{K-1} x_{k+1}, \quad \bar{y}_K = \frac{1}{K}\sum_{k=0}^{K-1} y_{k+1}.$$

Finally we derive for all $K \geq 1$

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{K}\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right).$$

□

3.3.2 Convex-strongly concave setting

For the remainder of this section we assume that the function g is convex with modulus $\nu > 0$, meaning it is ν-strongly convex. In this case the choice (26) leads to adaptive parameters and accelerated convergence.

Proposition 10. *Let $c_\alpha > L_{yx} \geq 0$, $\theta_0 = 1$ and $\tau_0, \sigma_0 > 0$ such that*

$$(c_\alpha L_{yx}\tau_0 + 2L_{yy})\,\sigma_0 < 1.$$

If $\nu > 0$ then $(\tau_k)_{k\geq 0}$, $(\sigma_k)_{k\geq 0}$ and $(\theta_k)_{k\geq 0}$ as defined in Proposition 6 are adaptive, in particular we have

$$\theta_{k+1} = \frac{1}{\sqrt{1+\nu\sigma_k}} < 1, \quad \tau_{k+1} = \frac{\tau_k}{\theta_{k+1}} > \tau_k, \quad \sigma_{k+1} = \theta_{k+1}\sigma_k < \sigma_k \quad \text{for all } k \geq 0. \tag{37}$$

Proof. The statements follow directly from Proposition 6 for $\nu > 0$. □

To obtain statements regarding the (accelerated) convergence rates in the convex-strongly concave setting, we look at the behaviour of the sequences of step size parameters $(\tau_k)_{k\geq 0}$ and $(\sigma_k)_{k\geq 0}$ for $k \to +\infty$.

Proposition 11. *Let $\theta_0 = 1$, $\tau_0 > 0$,*

$$0 < \sigma_0 \leq \frac{9 + 3\sqrt{13}}{2\nu},$$

and for all $k \geq 0$ denote

$$\gamma_k := \frac{\tau_k}{\sigma_k}.$$

Then with the choice of adaptive parameters (37) we have for all $k \geq 0$

$$\gamma_k \geq \frac{\nu^2\tau_0\sigma_0}{9}k^2 \quad \text{and} \quad \tau_k \geq \frac{\nu\tau_0\sigma_0}{3}k,$$

and for all $k \geq 1$

$$\sigma_k \leq \frac{3}{\nu}\frac{1}{k}.$$

Proof. By (26) we conclude that for all $k \geq 0$

$$\gamma_{k+1} = \gamma_k(1 + \nu\sigma_k),$$

and further

$$\sigma_{k+1} = \sigma_k \sqrt{\frac{\gamma_k}{\gamma_{k+1}}},$$

which, applied recursively, gives

$$\sigma_k = \sigma_0 \sqrt{\frac{\gamma_0}{\gamma_k}} = \sqrt{\tau_0\sigma_0}\frac{1}{\sqrt{\gamma_k}}.$$

We obtain

$$\gamma_{k+1} = \gamma_k(1 + \nu\sigma_k) = \gamma_k + \nu\sqrt{\tau_0\sigma_0}\sqrt{\gamma_k},$$

which we will use to show by induction that for all $k \geq 0$

$$\gamma_k \geq \frac{\nu^2\tau_0\sigma_0}{9}k^2. \tag{38}$$

For $k = 0$ the statement trivially holds, whereas for $k = 1$ we need to verify that

$$\gamma_1 = \gamma_0 + \nu\sqrt{\tau_0\sigma_0}\sqrt{\gamma_0} = \frac{\tau_0}{\sigma_0}(1 + \nu\sigma_0) \geq \frac{\nu^2\tau_0\sigma_0}{9},$$

which is equivalent to the following quadratic inequality

$$\sigma_0^2 - \frac{9}{\nu}\sigma_0 - \frac{9}{\nu^2} \leq 0,$$

and guaranteed to hold by our initial choice of $\sigma_0 > 0$. Now let $k \geq 1$ and assume that (38) holds. Then

$$\gamma_{k+1} = \gamma_k + \nu\sqrt{\tau_0\sigma_0}\sqrt{\gamma_k} \geq \frac{\nu^2\tau_0\sigma_0}{9}k^2 + \frac{\nu^2\tau_0\sigma_0}{3}k \geq \frac{\nu^2\tau_0\sigma_0}{9}(k+1)^2.$$

This shows the validity of (38) for all $k \geq 0$.

Now we can use inequality (38) to deduce the convergence behaviour of the sequences $(\tau_k)_{k\geq 0}$ and $(\sigma_k)_{k\geq 0}$ for $k \to +\infty$. We get for all $k \geq 0$

$$\tau_k = \sigma_k\gamma_k = \sqrt{\tau_0\sigma_0}\sqrt{\gamma_k} \geq \frac{\nu\tau_0\sigma_0}{3}k, \tag{39}$$

which, combined with

$$\tau_k\sigma_k = \frac{\tau_k^2}{\gamma_k} = \tau_0\sigma_0,$$

gives for all $k \geq 1$

$$\sigma_k \leq \frac{3}{\nu}\frac{1}{k}.$$

□

Now we are ready to prove the convergence results in the convex-strongly concave setting.

Theorem 12. *Let $c_\alpha > L_{yx} \geq 0$, $\theta_0 = 1$ and τ_0, $\sigma_0 > 0$ such that*

$$(c_\alpha L_{yx}\tau_0 + 2L_{yy})\,\sigma_0 < 1 \quad \text{and} \quad 0 < \sigma_0 \leq \frac{9 + 3\sqrt{13}}{2\nu}.$$

Let $(x^*, y^*) \in \mathcal{H}\times\mathcal{G}$ be a saddle point of (1). Then for $(x_k, y_k)_{k\geq 0}$ being the sequence generated by OGAProx with the choice of adaptive parameters

$$\theta_{k+1} = \frac{1}{\sqrt{1 + \nu\sigma_k}} < 1, \qquad \tau_{k+1} = \frac{\tau_k}{\theta_{k+1}} > \tau_k, \qquad \sigma_{k+1} = \theta_{k+1}\sigma_k < \sigma_k \quad \text{for all } k \geq 0,$$

we have for all $K \geq 1$

$$\|y^* - y_K\| \leq \frac{c_1}{K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right)^{\frac{1}{2}},$$

with $c_1 := \sqrt{\frac{18}{\nu^2\sigma_0\delta}}$*, where* $\delta > 0$ *is defined in* (28)*. Furthermore, for* $K \geq 1$*, denote*

$$T_K = \sum_{k=0}^{K-1} t_k, \quad \bar{x}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k x_{k+1}, \quad \bar{y}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k y_{k+1},$$

where $t_k = \frac{\tau_k}{\tau_0}$ *for all* $k \geq 0$ *(see also* (29)*). Then for all* $K \geq 2$ *it holds*

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{c_2}{K^2}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right),$$

with $c_2 := \frac{12}{\nu\sigma_0}$.

Proof. Let $K \geq 1$ and let $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ be an arbitrary but fixed saddle point. First we will prove the convergence rate of the sequence of iterates $(y_k)_{k\geq 0}$. Plugging the particular choice of parameters (37) into (24) for (x^*, y^*), we obtain

$$\begin{aligned}\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2 &\geq \frac{1}{2\tau_0}\|x^* - x_K\|^2 + \frac{\tau_K}{\sigma_K}\left(1 - \sigma_K\theta_K(L_{yx}\alpha_K + L_{yy})\right)\frac{1}{2\tau_0}\|y^* - y_K\|^2 \\ &\geq \gamma_K\frac{\delta}{2\tau_0}\|y^* - y_K\|^2,\end{aligned}$$

where we use (10) in Assumption 1 for the last inequality. Combining this with (38) we derive

$$\|y^* - y_K\| \leq \frac{c_1}{K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right)^{\frac{1}{2}},$$

with $c_1 := \sqrt{\frac{18}{\nu^2\sigma_0\delta}}$.

Next we will show the convergence rate of the minimax gap at the ergodic sequences. Writing (25) for (x^*, y^*), we obtain

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{T_K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right). \tag{40}$$

Plugging the particular choice of $t_k = \frac{\tau_k}{\tau_0}$ for all $k \geq 0$ from (29) into the definition of T_K, together with (39) yields

$$T_K = \frac{1}{\tau_0}\sum_{k=0}^{K-1}\tau_k \geq \frac{\nu\sigma_0}{3}\sum_{k=0}^{K-1} k = \frac{\nu\sigma_0}{6}K(K-1).$$

Combining this inequality with (40), we obtain for all $K \geq 2$

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{c_2}{K^2}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right),$$

with $c_2 := \frac{12}{\nu\sigma_0}$, which concludes the proof. □

4 Strongly convex-strongly concave setting

For this section we assume that the function g is convex with modulus $\nu > 0$, meaning it is ν-strongly convex. In addition to the assumptions we had until now, for this section we also assume that for all $y \in \operatorname{dom} g$ the function $\Phi(\cdot\,, y) : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is μ-strongly convex with modulus $\mu > 0$. This means that the saddle function $(x, y) \mapsto \Psi(x, y)$ is strongly convex-strongly concave.

As in the previous section we will state two step size assumptions that will be needed for the convergence analysis. These again will be followed by preparatory observations and a result to guarantee the validity of the stated assumptions. The section will be closed with the formulation and proof of convergence results.

Assumption 2. *We assume that the step sizes τ_k, σ_k and the momentum parameter θ_k are constant*

$$\theta_k = \theta_0 =: \theta, \quad \tau_k = \tau_0 =: \tau, \quad \sigma_k = \sigma_0 =: \sigma \quad \forall k \geq 0,$$

and satisfy

$$1 + \mu\tau = \frac{1}{\theta}, \quad 1 + \nu\sigma = \frac{1}{\theta}, \tag{41}$$

with

$$0 < \theta < 1. \tag{42}$$

Furthermore, we assume that there exists $\alpha > 0$ such that

$$\frac{L_{yx}}{\alpha} \leq \frac{1}{\tau}, \quad L_{yy} \leq \frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{\sigma}, \tag{43}$$

with

$$1 - \theta\sigma(\alpha L_{yx} + L_{yy}) > 0. \tag{44}$$

4.1 Preliminary considerations

We take an arbitrary $(x, y) \in \mathcal{H} \times \mathcal{G}$ and let $k \geq 0$. Following similar considerations along (13)-(15), additionally taking into account the μ-strong convexity of $\Phi(\cdot\,, y)$ for $y \in \operatorname{dom} g$, instead of (16) we derive

$$\begin{aligned}
\Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) &\leq \theta \langle q_k, y_k - y \rangle - \langle q_{k+1}, y_{k+1} - y \rangle \\
&\quad - \frac{\mu}{2} \|x - x_{k+1}\|^2 - \frac{\nu}{2} \|y - y_{k+1}\|^2 + \theta \langle q_k, y_{k+1} - y_k \rangle \\
&\quad + \frac{1}{2\tau} \left(- \|x_k - x_{k+1}\|^2 - \|x - x_{k+1}\|^2 + \|x - x_k\|^2 \right) \\
&\quad + \frac{1}{2\sigma} \left(- \|y_k - y_{k+1}\|^2 - \|y - y_{k+1}\|^2 + \|y - y_k\|^2 \right) \\
&\leq \frac{1}{2\tau} \|x - x_k\|^2 + \frac{1}{2\sigma} \|y - y_k\|^2 + \theta \langle q_k, y_k - y \rangle \\
&\quad - \frac{1 + \mu\tau}{2\tau} \|x - x_{k+1}\|^2 - \frac{1 + \nu\sigma}{2\sigma} \|y - y_{k+1}\|^2 - \langle q_{k+1}, y_{k+1} - y \rangle \\
&\quad + \frac{\theta L_{yx}}{2\alpha} \|x_k - x_{k-1}\|^2 - \frac{1}{2\tau} \|x_{k+1} - x_k\|^2 \\
&\quad + \frac{\theta L_{yy}}{2} \|y_k - y_{k-1}\|^2 - \frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{2\sigma} \|y_{k+1} - y_k\|^2 .
\end{aligned}$$

By (41) in Assumption 2 and for $\alpha > 0$ fulfilling (43)-(44), we obtain

$$\begin{aligned}
\Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) &\leq \frac{1}{2\tau} \|x - x_k\|^2 + \frac{1}{2\sigma} \|y - y_k\|^2 + \theta \langle q_k, y_k - y \rangle \\
&\quad - \frac{1}{2\tau}\frac{1}{\theta} \|x - x_{k+1}\|^2 - \frac{1}{2\sigma}\frac{1}{\theta} \|y - y_{k+1}\|^2 - \langle q_{k+1}, y_{k+1} - y \rangle \\
&\quad + \frac{\theta L_{yx}}{2\alpha} \|x_k - x_{k-1}\|^2 - \frac{1}{2\tau} \|x_{k+1} - x_k\|^2 \\
&\quad + \frac{\theta L_{yy}}{2} \|y_k - y_{k-1}\|^2 - \frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{2\sigma} \|y_{k+1} - y_k\|^2 ,
\end{aligned}$$

which together with (43) and (44) gives

$$\begin{aligned}\Psi(x_{k+1},y)-\Psi(x,y_{k+1}) \le{}& \frac{1}{2\tau}\left(\|x-x_k\|^2-\frac{1}{\theta}\|x-x_{k+1}\|^2\right)+\frac{1}{2\sigma}\left(\|y-y_k\|^2-\frac{1}{\theta}\|y-y_{k+1}\|^2\right)\\ &+\theta\langle q_k,y_k-y\rangle-\langle q_{k+1},y_{k+1}-y\rangle+\frac{1}{2\tau}\left(\theta\|x_k-x_{k-1}\|^2-\|x_{k+1}-x_k\|^2\right)\\ &+\frac{1}{2\tilde{\sigma}}\left(\theta\|y_k-y_{k-1}\|^2-\|y_{k+1}-y_k\|^2\right),\end{aligned} \tag{45}$$

where

$$\tilde{\sigma}:=\frac{\sigma}{1-\theta\sigma(\alpha L_{yx}+L_{yy})}.$$

Let $K\ge 1$ and as in (21) denote

$$T_K=\sum_{k=0}^{K-1}t_k,\qquad \bar{x}_K=\frac{1}{T_K}\sum_{k=0}^{K-1}t_kx_{k+1},\qquad \bar{y}_K=\frac{1}{T_K}\sum_{k=0}^{K-1}t_ky_{k+1}.$$

with $t_k>0$ defined as in (19), in other words

$$t_k=\theta^{-k}\quad\forall k\ge 0.$$

Multiplying both sides of (45) by $t_k>0$ yields

$$\begin{aligned}\frac{1}{\theta^k}\left(\Psi(x_{k+1},y)-\Psi(x,y_{k+1})\right) \le{}& \frac{1}{2\tau}\left(\frac{1}{\theta^k}\|x-x_k\|^2-\frac{1}{\theta^{k+1}}\|x-x_{k+1}\|^2\right)\\ &+\frac{1}{2\sigma}\left(\frac{1}{\theta^k}\|y-y_k\|^2-\frac{1}{\theta^{k+1}}\|y-y_{k+1}\|^2\right)\\ &+\frac{1}{\theta^{k-1}}\langle q_k,y_k-y\rangle-\frac{1}{\theta^k}\langle q_{k+1},y_{k+1}-y\rangle\\ &+\frac{1}{2\tau}\left(\frac{1}{\theta^{k-1}}\|x_k-x_{k-1}\|^2-\frac{1}{\theta^k}\|x_{k+1}-x_k\|^2\right)\\ &+\frac{1}{2\tilde{\sigma}}\left(\frac{1}{\theta^{k-1}}\|y_k-y_{k-1}\|^2-\frac{1}{\theta^k}\|y_{k+1}-y_k\|^2\right).\end{aligned}$$

Summing up the above inequality for $k=0,\dots,K-1$ and taking into account Jensen's inequality for the

convex function $\Psi(\,\cdot\,, y) - \Psi(x, \,\cdot\,)$ give

$$\begin{aligned}
T_K\left(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)\right) &\leq \sum_{k=0}^{K-1} \frac{1}{\theta^k}\left(\Psi(x_{k+1}, y) - \Psi(x, y_{k+1})\right) \\
&\leq \frac{1}{2\tau}\left(\|x - x_0\|^2 - \frac{1}{\theta^K}\|x - x_K\|^2\right) + \frac{1}{2\sigma}\left(\|y - y_0\|^2 - \frac{1}{\theta^K}\|y - y_K\|^2\right) \\
&\quad - \frac{1}{\theta^{K-1}}\langle q_K, y_K - y\rangle - \frac{1}{\theta^{K-1}}\frac{1}{2\tau}\|x_K - x_{K-1}\|^2 - \frac{1}{\theta^{K-1}}\frac{1}{2\tilde{\sigma}}\|y_K - y_{K-1}\|^2 \\
&\leq \frac{1}{2\tau}\left(\|x - x_0\|^2 - \frac{1}{\theta^K}\|x - x_K\|^2\right) + \frac{1}{2\sigma}\left(\|y - y_0\|^2 - \frac{1}{\theta^K}\|y - y_K\|^2\right) \\
&\quad + \frac{1}{\theta^{K-1}}\frac{L_{yx}}{2}\left(\frac{1}{\alpha}\|x_K - x_{K-1}\|^2 + \alpha\|y_K - y\|^2\right) - \frac{1}{\theta^{K-1}}\frac{1}{2\tau}\|x_K - x_{K-1}\|^2 \\
&\quad + \frac{1}{\theta^{K-1}}\frac{L_{yy}}{2}\left(\|y_K - y_{K-1}\|^2 + \|y_K - y\|^2\right) - \frac{1}{\theta^{K-1}}\frac{1}{2\tilde{\sigma}}\|y_K - y_{K-1}\|^2 \\
&= \frac{1}{2\tau}\|x - x_0\|^2 + \frac{1}{2\sigma}\|y - y_0\|^2 \\
&\quad - \frac{1}{\theta^K}\frac{1}{2\tau}\|x - x_K\|^2 - \frac{1}{\theta^K}\frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{2\sigma}\|y - y_K\|^2 \\
&\quad - \frac{1}{2\theta^{K-1}}\left(\frac{1}{\tau} - \frac{L_{yx}}{\alpha}\right)\|x_K - x_{K-1}\|^2 - \frac{1}{2\theta^{K-1}}\left(\frac{1}{\tilde{\sigma}} - L_{yy}\right)\|y_K - y_{K-1}\|^2,
\end{aligned}$$

where in the second inequality we use (17). Omitting the last two terms which are non positive by (43), we obtain for all $K \geq 1$

$$T_K\theta^K\left(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)\right) + \frac{1}{2\tau}\|x - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y - y_K\|^2 \leq \theta^K\left(\frac{1}{2\tau}\|x - x_0\|^2 + \frac{1}{2\sigma}\|y - y_0\|^2\right), \tag{46}$$

which we will use to obtain our convergence results in the following.

4.2 Fulfilment of step size assumptions

In this subsection we will investigate a particular choice of parameters τ, σ and θ such that Assumption 2 holds.

Proposition 13. *For $\alpha > 0$ define*

$$\tilde{\theta} := \max\left\{\frac{L_{yx}}{\alpha\mu + L_{yx}}, \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}}\right\}. \tag{47}$$

Let $\theta > 0$ such that

$$0 \leq \tilde{\theta} < \theta < 1, \tag{48}$$

and set

$$\tau = \frac{1}{\mu}\frac{1 - \theta}{\theta} \quad \text{and} \quad \sigma = \frac{1}{\nu}\frac{1 - \theta}{\theta}. \tag{49}$$

Then τ, σ and θ fulfil Assumption 2.

Proof. If $L_{yx} = L_{yy} = 0$, then the conclusion follows immediately. Assume that $L_{yx} + L_{yy} > 0$. It is easy to verify that definition (47) yields

$$0 < \tilde{\theta} < 1$$

and that (49) is equivalent to (41) where (42) is ensured by (48). Furthermore, plugging the specific form of the step sizes (49) into (43) we obtain for the first inequality of (43)

$$\frac{L_{yx}}{\alpha} \leq \frac{\mu\theta}{1-\theta},$$

which is equivalent to

$$\theta \geq \frac{L_{yx}}{\alpha\mu + L_{yx}}.$$

Note that by (48) we have

$$0 \leq \frac{L_{yx}}{\alpha\mu + L_{yx}} \leq \tilde{\theta} < \theta < 1.$$

Similarly, the second inequality of (43) is equivalent to the following quadratic inequality

$$\theta^2 - \frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}}\theta - \frac{L_{yy}}{\alpha L_{yx} + L_{yy}} \geq 0.$$

The non negative solution of the associated quadratic equation reads

$$\rho := \frac{1}{2}\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}} + \sqrt{\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}}\right)^2 + \frac{4L_{yy}}{\alpha L_{yx} + L_{yy}}}\right) \geq 0.$$

Since

$$0 \leq \rho < \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}} \leq \tilde{\theta} < \theta < 1,$$

the second inequality of (43) is also fulfilled. In order to see that

$$\rho < \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}},$$

we notice that this inequality is equivalent to

$$\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}}\right)^2 + \frac{4L_{yy}}{\alpha L_{yx} + L_{yy}} < \frac{\left(\nu^2 + 2\nu L_{yy} + (\alpha L_{yx} + 2L_{yy})^2\right)^2}{(\nu + \alpha L_{yx} + 2L_{yy})^2(\alpha L_{yx} + L_{yy})^2},$$

which holds if and only if

$$\begin{aligned}&\big((\alpha L_{yx} - \nu)^2 + 4L_{yy}(\alpha L_{yx} + L_{yy})\big)(\nu + \alpha L_{yx} + 2L_{yy})^2 \\ &< (\nu^2 + 2\nu L_{yy})^2 + 2(\nu^2 + 2\nu L_{yy})(\alpha L_{yx} + 2L_{yy})^2 + (\alpha L_{yx} + 2L_{yy})^4\end{aligned}$$

or, equivalently,

$$0 < 4\nu^2(\alpha L_{yx} + L_{yy})^2.$$

For the remaining condition (44) to hold we need to ensure

$$\theta > \frac{\alpha L_{yx} + L_{yy} - \nu}{\alpha L_{yx} + L_{yy}}.$$

For this we observe that

$$\begin{aligned}\rho &\geq \frac{1}{2}\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}} + \sqrt{\left(\frac{\alpha L_{yx} + 2L_{yy} - \nu}{\alpha L_{yx} + L_{yy}}\right)^2}\right) \\ &\geq \frac{1}{2}\frac{\alpha L_{yx} - \nu + \alpha L_{yx} + 2L_{yy} - \nu}{\alpha L_{yx} + L_{yy}} = \frac{\alpha L_{yx} + L_{yy} - \nu}{\alpha L_{yx} + L_{yy}}.\end{aligned}$$

In conclusion, we obtain the following chain of inequalities

$$\frac{\alpha L_{yx} + L_{yy} - \nu}{\alpha L_{yx} + L_{yy}} \leq \rho < \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}} \leq \tilde{\theta} < \theta < 1,$$

which is satisfied by (48). □

4.3 Convergence results

Now we can combine the previous results and prove the convergence statements in the strongly convex-strongly concave setting.

Theorem 14. *Let $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ be a saddle point of (1). Then for $(x_k, y_k)_{k\geq 0}$ being the sequence generated by OGAProx with the choice of parameters*

$$\tau = \frac{1}{\mu}\frac{1-\theta}{\theta}, \quad \sigma = \frac{1}{\nu}\frac{1-\theta}{\theta}, \quad 0 \leq \tilde{\theta} < \theta < 1,$$

with

$$\tilde{\theta} = \max\left\{\frac{L_{yx}}{\alpha\mu + L_{yx}}, \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}}\right\},$$

for $\alpha > 0$, we denote for $K \geq 1$

$$T_K = \sum_{k=0}^{K-1} \theta^{-k}, \quad \bar{x}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} \theta^{-k} x_{k+1}, \quad \bar{y}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} \theta^{-k} y_{k+1},$$

for which the following holds

$$0 \leq \theta\big(\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)\big) + \frac{1}{2\tau}\|x^* - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y^* - y_K\|^2 \leq \theta^K\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right),$$

where $\tilde{\sigma} := \frac{\sigma}{1-\theta\sigma(\alpha L_{yx} + L_{yy})}$.

Proof. Let $K \geq 1$ and $(x^*, y^*) \in \mathcal{H} \times \mathcal{G}$ be an arbitrary but fixed saddle point of (1). Writing (46) for (x^*, y^*) we get

$$\begin{aligned} 0 &\leq T_K\theta^K\big(\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)\big) + \frac{1}{2\tau}\|x^* - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y^* - y_K\|^2 \\ &\leq \theta^K\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right). \end{aligned}$$

Using

$$T_K = \sum_{k=0}^{K-1}\frac{1}{\theta^k} = \frac{1}{\theta^{K-1}}\frac{1-\theta^K}{1-\theta} \geq \frac{1}{\theta^{K-1}},$$

finally we obtain for all $K \geq 1$

$$0 \leq \theta\big(\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)\big) + \frac{1}{2\tau}\|x^* - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y^* - y_K\|^2 \leq \theta^K\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right),$$

with $0 < \theta < 1$ as defined in (48). □

5 Numerical experiments

In this section we will treat three numerical applications of our method. The first one is of rather simple structure and has the purpose to highlight the convergence rates we obtained in the previous sections. The second one concerns multi kernel support vector machines to validate OGAProx on a more relevant application in practice, even though there are no theoretical guarantees for the "metric" reported there. The third numerical application addresses a classification problem incorporating minimax group fairness, which traces back to the solving of a minimax problem with nonsmooth coupling function.

5.1 Nonsmooth-linear problem

The first application we treat is to showcase the convergence rates we obtained in the previous sections and make a simple proof of concept. We look at the following nonsmooth-linear saddle point problem

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n}\Psi(x,y) := \langle [x]_+, Ay\rangle - \left(\delta_C(y) + \frac{\nu}{2}\|y\|^2\right), \tag{50}$$

with $\nu \geq 0$ and $A \in \mathbb{R}^{d\times n}$, $[\,\cdot\,]_+$ being the component-wise positive part,

$$[x]_+ = \big(\max\{0, x_i\}\big)_{i=1}^d,$$

and C being the following convex polytope

$$C := \{y \in \mathbb{R}^n \mid Ay \geqq 0\}.$$

For $u = (u_i)_{i=1}^d$, $v = (v_i)_{i=1}^d \in \mathbb{R}^d$ the relation $u \geqq v$ denotes component-wise inequalities, namely,

$$u \geqq v \;\Leftrightarrow\; u_i \geq v_i \quad \text{for } 1 \leq i \leq d.$$

Then $g : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ with

$$g(y) := \delta_C(y) + \frac{\nu}{2}\|y\|^2$$

is proper, lower semicontinuous and convex with modulus $\nu \geq 0$ and $\operatorname{dom} g = C$. Moreover, $\Phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ with

$$\Phi(x,y) = \sum_{i=1}^d \max\{0, x_i\}(Ay)_i$$

has full domain, for all $x \in \mathbb{R}^d$ we have that $\Phi(x,\,\cdot\,)$ is linear and for all $y \in \operatorname{dom} g = C$ the function $\Phi(\,\cdot\,, y)$ is convex and continuous.

Furthermore, we obtain for all (x,y), $(x',y') \in \mathbb{R}^d \times \operatorname{dom} g$

$$\left\|\nabla_y\Phi(x,y) - \nabla_y\Phi(x',y')\right\| = \left\|A^T\left([x]_+ - [x']_+\right)\right\| \leq \|A\|\left\|x - x'\right\|,$$

hence (2) holds with $L_{yx} = \|A\|$ and $L_{yy} = 0$.

The algorithm (7)-(8) iterates for $k \geq 0$

$$\begin{cases} v_k &= y_k + \sigma_k\left[(1+\theta_k)\nabla_y\Phi(x_k,y_k) - \theta_k\nabla_y\Phi(x_{k-1},y_{k-1})\right] = y_k + \sigma_k A^T\big((1+\theta_k)[x_k]_+ - \theta_k[x_{k-1}]_+\big), \\ y_{k+1} &= \operatorname{prox}_{\sigma_k g}(v_k) = P_C\left(\frac{1}{1+\nu\sigma_k}v_k\right), \\ x_{k+1} &= \operatorname{prox}_{\tau_k\Phi(\,\cdot\,,y_{k+1})}(x_k), \end{cases}$$

where the calculation of the orthogonal projection on the set C is a simple quadratic program and

$$\operatorname{prox}_{\tau\Phi(\,\cdot\,,y)}(x) = \Big(\operatorname{prox}_{\tau(Ay)_i\max\{0,\,\cdot\,\}}(x_i)\Big)_{i=1}^d,$$

where, for $i = 1, ..., d$,

$$\operatorname{prox}_{\tau(Ay)_i\max\{0,\,\cdot\,\}}(x_i) = \begin{cases} x_i & \text{if } x_i \leq 0, \\ 0 & \text{if } 0 < x_i \leq \tau(Ay)_i, \\ x_i - \tau(Ay)_i & \text{if } x_i > \tau(Ay)_i. \end{cases}$$

By writing the first order optimality conditions and using Lagrange duality we obtain the following characterisation.

$$
\begin{aligned}
(x^*, y^*) \text{ is a saddle point of (50)} \Leftrightarrow & \begin{cases} 0 \in \partial \left(\langle [\,\cdot\,]_+, Ay^* \rangle - \delta_C(y^*) - \frac{\nu}{2} \|y^*\|^2 \right)(x^*) \\ 0 \in \partial \left(- \left\langle A^T[x^*]_+, \cdot \right\rangle + \delta_C(\,\cdot\,) + \frac{\nu}{2} \|\,\cdot\,\|^2 \right)(y^*) \end{cases} \\
\Leftrightarrow & \begin{cases} 0 \in \sum_{i=1}^d (Ay^*)_i \, \partial \max\{0, \cdot\}(x_i^*) \\ A^T[x^*]_+ - \nu y^* \in N_C(y^*) \end{cases} \\
\Leftrightarrow & \begin{cases} \forall i = 1, \dots, d : \big((Ay^*)_i > 0 \text{ and } x_i^* \leq 0 \big) \text{ or} \\ \quad \big((Ay^*)_i = 0 \text{ and } x_i^* \in \mathbb{R} \big) \\ \left\langle A^T[x^*]_+ - \nu y^*, y^* \right\rangle = 0 \\ \nu y^* - A^T[x^*]_+ \in A^T \left(\mathbb{R}^d_+ \right) \end{cases} \\
\Leftrightarrow & \begin{cases} \forall i = 1, \dots, d : \big((Ay^*)_i > 0 \text{ and } x_i^* \leq 0 \big) \text{ or} \\ \quad \big((Ay^*)_i = 0 \text{ and } x_i^* \in \mathbb{R} \big) \\ \nu \|y^*\|^2 = \left\langle A^T[x^*]_+, y^* \right\rangle = \langle [x^*]_+, Ay^* \rangle = 0 \\ \nu y^* \in A^T \left([x^*]_+ + \mathbb{R}^d_+ \right). \end{cases}
\end{aligned}
$$

This means, that for $\nu = 0$ we obtain

$$
(x^*, y^*) \text{ is a saddle point of (50)} \Leftrightarrow \begin{cases} \forall i = 1, \dots, d : \big((Ay^*)_i > 0 \text{ and } x_i^* \leq 0 \big) \text{ or} \\ \quad \big((Ay^*)_i = 0 \text{ and } x_i^* \in \mathbb{R} \big) \\ 0 \in A^T \left([x^*]_+ + \mathbb{R}^d_+ \right) \end{cases},
$$

whereas for $\nu > 0$

$$
(x^*, y^*) \text{ is a saddle point of (50)} \Leftrightarrow \begin{cases} y^* = 0 \\ 0 \in A^T \left([x^*]_+ + \mathbb{R}^d_+ \right) \end{cases}.
$$

If $A \in \mathbb{R}^{d \times n}$ has full row rank the inclusion

$$
0 \in A^T \left([x^*]_+ + \mathbb{R}^d_+ \right)
$$

is equivalent to

$$
x^* \leqq 0.
$$

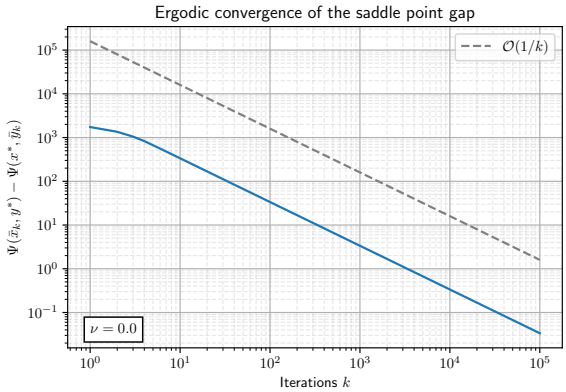

Figure 1: Convergence of the minimax gap like $\mathcal{O}(\frac{1}{K})$ for $\nu = 0$.

For the experiments we choose dimensions $d = 250$ and $n = 350$. For easier validation of the solution x^* we ensure that the matrix $A \in \mathbb{R}^{d \times n}$ with entries drawn from a uniform distribution on the interval

$[-3, 3]$ has full row rank. The starting points $x_0 = x_{-1} \in \mathbb{R}^d$ and $y_0 \in \mathbb{R}^n$ have entries drawn from a uniform distribution on the interval $[-5, 5]$.

In the case $\nu = 0$, i.e., the regulariser g being merely convex, we proved weak asymptotic convergence of the iterates to some saddle point (x^*, y^*) and convergence of the minimax gap at the ergodic sequences to zero like $\mathcal{O}(\frac{1}{K})$ for any saddle point. The latter is illustrated in Figure 1 for $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^n$ with $x^* \leqq 0$ and $y^* \in C$ with $y^* \neq 0$ for a single random initialisation.

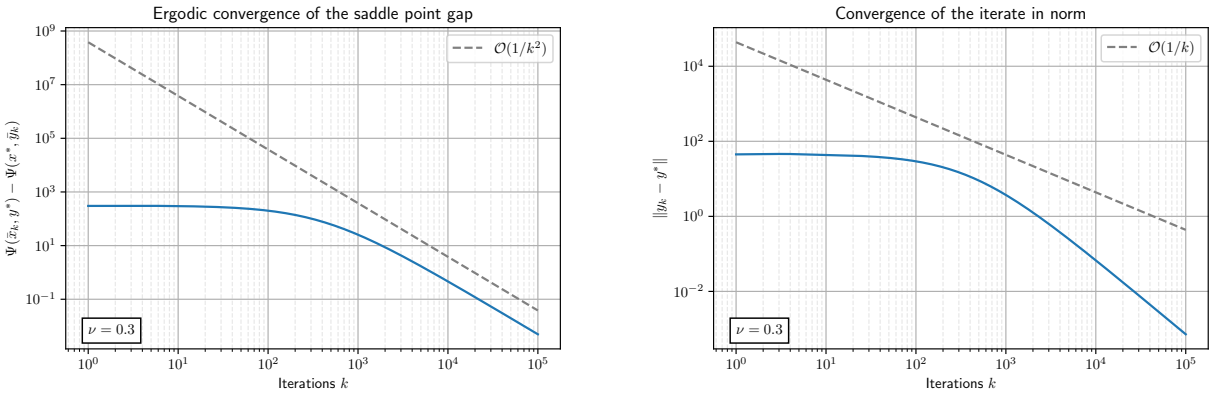

Figure 2: Convergence of the minimax gap like $\mathcal{O}(\frac{1}{K^2})$ and of the sequence $(y_k)_{k\geq 0}$ in norm like $\mathcal{O}(\frac{1}{K})$ for $\nu > 0$.

Let $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^n$ be a saddle point. In the case $\nu > 0$, i.e., the regulariser g being ν-strongly convex, we proved strong non-asymptotic convergence of the sequence $(y_k)_{k\geq 0} \to y^*$ like $\mathcal{O}(\frac{1}{K})$ and convergence of the minimax gap at the ergodic sequences to zero like $\mathcal{O}(\frac{1}{K^2})$. The numerical behaviour of our method validating the theoretical claims for $\nu > 0$ is highlighted in Figure 2. The plots shown are for a single random initialisation and with the choice $\nu = \frac{3}{10}$.

5.2 Multi kernel support vector machine

The second application to test our method in practice is to learn a combined kernel matrix for a multi kernel *support vector machine* (SVM). We have a set of labelled training data

$$S_n = \{(a_1, b_1), \ldots, (a_n, b_n)\} \subseteq \mathbb{R}^m \times \{-1, 1\},$$

where we call $b = (b_i)_{i=1}^n$, and a set of unlabelled test data

$$T_l = \{a_{n+1}, \ldots, a_{n+l}\} \subseteq \mathbb{R}^m.$$

We consider embeddings of the data according to a kernel function $\kappa : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ with the corresponding symmetric and positive semidefinite kernel matrix

$$\mathcal{K} = \begin{pmatrix} \mathcal{K}^{tr} & \mathcal{K}^{tr,t} \\ \mathcal{K}^{t,tr} & \mathcal{K}^{t} \end{pmatrix},$$

where $\mathcal{K}_{ij} = \kappa(a_i, a_j)$ for $i, j = 1, \ldots, n, n+1, \ldots, n+l$.

In the following e is a vector of appropriate size consisting of ones. According to [13] the problem of interest is

$$\min_{\substack{\mathcal{K}\in\mathbb{K} \\ \operatorname{trace}(\mathcal{K})=c}} \max_{\substack{0\leqq\alpha\leqq C \\ \langle\alpha,b\rangle=0}} \alpha^T e - \frac{1}{2}\alpha^T G(\mathcal{K}^{tr})\alpha - \frac{\nu}{2}\|\alpha\|_2^2, \tag{51}$$

where $\mathbb{K}$ is the model class of kernel matrices, $c \in (0, +\infty)$, $C \in (0, +\infty]$ and $\nu \in [0, +\infty)$ are model parameters and we define $G(\mathcal{K}^{tr}) := \operatorname{diag}(b)\mathcal{K}^{tr}\operatorname{diag}(b)$.

The set $\mathbb{K}$ is restricted to be the set of positive semidefinite matrices that can be written as a non negative linear combination of kernel matrices $\mathcal{K}_1, \ldots, \mathcal{K}_d$, i.e.,

$$\mathbb{K} = \left\{ \mathcal{K} \in S_+^m \;\middle|\; \mathcal{K} = \sum_{i=1}^{d} \eta_i \mathcal{K}_i,\ \eta_i \geq 0 \text{ for } i = 1, ..., d \right\}.$$

With this choice (51) becomes

$$\min_{\substack{\langle \eta, r \rangle = c \\ \eta \geqq 0}} \max_{\substack{0 \leqq \alpha \leqq C \\ \langle \alpha, b \rangle = 0}} \alpha^T e - \frac{1}{2} \sum_{i=1}^{d} \eta_i \alpha^T G(\mathcal{K}_i^{tr}) \alpha - \frac{\nu}{2} \|\alpha\|^2, \tag{52}$$

where $\eta = (\eta_i)_{i=1}^d$ and $r = (r_i)_{i=1}^d$ with $r_i = \operatorname{trace}(\mathcal{K}_i)$ for $i = 1, ..., d$. Assume $(\eta^*, \alpha^*) \in \mathbb{R}^d \times \mathbb{R}^n$ to be a saddle point of (52) and write

$$\mathcal{K}^* = \sum_{j=1}^{d} \eta_j^* \mathcal{K}_j.$$

Following the considerations of [11] we compute for $a_k \in T_l$ with $k \in \{n+1, \ldots, n+l\}$,

$$\mathcal{L}(a_k) = \operatorname{sgn}\left(\sum_{i=1}^{n} b_i \alpha_i^* \mathcal{K}_{ik}^* + \gamma \right) = \operatorname{sgn}\left(\sum_{i=1}^{n} \sum_{j=1}^{d} b_i \alpha_i^* \eta_j^* \left(\mathcal{K}_j\right)_{ik} + \gamma \right), \tag{53}$$

with

$$\gamma = b_{j_0}(1 - \nu \alpha_{j_0}^*) - \sum_{i=1}^{n} b_i \alpha_i^* \mathcal{K}_{ij_0}^* = b_{j_0}(1 - \nu \alpha_{j_0}^*) - \sum_{i=1}^{n} \sum_{j=1}^{d} b_i \alpha_i^* \eta_j^* \left(\mathcal{K}_j\right)_{ij_0},$$

for some $j_0 \in \{1, \ldots, n\}$ such that $0 < \alpha_{j_0}^* < C$.

After writing $x_i = \frac{r_i \eta_i}{c}$ for $i = 1, ..., d$ and augmenting the objective with an additional (strongly) convex penalisation term, we obtain

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^n} \delta_\Delta(x) + \frac{\mu}{2} \|x\|^2 - \frac{1}{2} \sum_{i=1}^{d} x_i y^T M_i y + y^T e - \left(\delta_Y(y) + \frac{\nu}{2} \|y\|^2 \right), \tag{54}$$

where $\mu \geq 0$ and $M_i := \frac{c}{r_i} G(\mathcal{K}_i^{tr})$ for $i = 1, ..., d$,

$$\Delta := \{ x \in \mathbb{R}^d \mid x \geqq 0,\ \langle x, e \rangle = 1 \}$$

is the m-dimensional unit simplex and

$$Y := \{ y \in \mathbb{R}^n \mid 0 \leqq y \leqq C,\ \langle y, b \rangle = 0 \}$$

is the intersection of a box and a hyperplane.

In the notation of (1) we have $\Phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\Phi(x, y) = \delta_\Delta(x) + \frac{\mu}{2} \|x\|^2 - \frac{1}{2} \sum_{i=1}^{d} x_i y^T M_i y + y^T e,$$

and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ given by

$$g(y) = \delta_Y(y) + \frac{\nu}{2} \|y\|^2.$$

We see that Φ and g satisfy the assumptions considered for problem (1).

The algorithm (7)-(8) iterates as follows for $k \geq 0$

$$\begin{cases} v_k &= y_k + \sigma_k \left[(1+\theta_k) \nabla_y \Phi(x_k, y_k) - \theta_k \nabla_y \Phi(x_{k-1}, y_{k-1}) \right], \\ y_{k+1} &= \operatorname{prox}_{\sigma_k g}(v_k) = P_Y\left(\frac{1}{1+\nu\sigma_k} v_k \right), \\ x_{k+1} &= \operatorname{prox}_{\tau_k \Phi(\,\cdot\,, y_{k+1})}(x_k) = P_\Delta\left(\frac{1}{1+\mu\tau_k}(x_k + \tau_k \xi^{y_{k+1}}) \right), \end{cases}$$

where

$$\nabla_y \Phi(x,y) = -\left(\frac{1}{2} \sum_{i=1}^d x_i (M_i + M_i^T) \right) y + e = -\left(\sum_{i=1}^d x_i M_i \right) y + e \quad \text{for } (x,y) \in \Delta \times \mathbb{R}^n$$

and

$$\xi^y := \left(\frac{1}{2} y^T M_i y \right)_{i=1}^d .$$

To determine the correct step sizes and momentum parameter, we need to find Lipschitz constants for $\nabla_y \Phi$, i.e., L_{yx}, $L_{yy} \geq 0$ such that (2) holds. Recall, that we require for all (x,y), $(x',y') \in \Pr_{\mathcal{H}}(\operatorname{dom} \Phi) \times \operatorname{dom} g$

$$\left\| \nabla_y \Phi(x,y) - \nabla_y \Phi(x',y') \right\| \leq L_{yx} \left\| x - x' \right\| + L_{yy} \left\| y - y' \right\|,$$

with $\Pr_{\mathcal{H}}(\operatorname{dom} \Phi) = \Delta$ and $\operatorname{dom} g = Y$.

Let (x,y), $(x',y') \in \Delta \times Y$. Then

$$\begin{aligned} \left\| \nabla_y \Phi(x,y) - \nabla_y \Phi(x',y') \right\| &= \left\| -\sum_{i=1}^d x_i M_i y + e + \sum_{i=1}^d x_i' M_i y' - e \right\| \\ &= \left\| \sum_{i=1}^d x_i M_i y' - \sum_{i=1}^d x_i M_i y + \sum_{i=1}^d x_i' M_i y' - \sum_{i=1}^d x_i M_i y' \right\| \\ &\leq \left\| \sum_{i=1}^d x_i M_i (y - y') \right\| + \left\| \sum_{i=1}^d (x_i - x_i') M_i y' \right\| \\ &\leq \sum_{i=1}^d |x_i| \left\| M_i \right\| \left\| y - y' \right\| + \sum_{i=1}^d |x_i - x_i'| \left\| M_i \right\| \left\| y' \right\| \\ &\leq \left(\|x\|_1 \max_{1 \leq i \leq d} \|M_i\| \right) \left\| y - y' \right\| + \left(\left\| y' \right\| \max_{1 \leq i \leq d} \|M_i\| \right) \left\| x - x' \right\|_1 \\ &\leq \left(\|x\|_1 \max_{1 \leq i \leq d} \|M_i\| \right) \left\| y - y' \right\| + \left(\left\| y' \right\| \sqrt{d} \max_{1 \leq i \leq d} \|M_i\| \right) \left\| x - x' \right\|. \end{aligned}$$

As $x \in \Delta$, we have $\|x\|_1 = 1$ and since $y' \in Y$ we get $\|y'\| \leq C\sqrt{n}$. Thus we obtain

$$\left\| \nabla_y \Phi(x,y) - \nabla_y \Phi(x',y') \right\| \leq L_{yx} \left\| x - x' \right\| + L_{yy} \left\| y - y' \right\|,$$

with

$$L_{yx} = C\sqrt{dn} \max_{1 \leq i \leq d} \|M_i\|, \qquad L_{yy} = \max_{1 \leq i \leq d} \|M_i\|.$$

For our experiments we use four different data sets from the "UCI Machine Learning Repository" [8]: the (original) Wisconsin *breast cancer* dataset [16] (699 total observations including 16 incomplete examples; 9 features), the Statlog *heart disease* data set (270 observations; 13 features), the *Ionosphere* data set (351 observations; 33 features) and the Connectionist Bench *Sonar* data set (208 observations; 60 features). All the data sets are normalised such that each feature column has zero mean and standard deviation equal to one.

Furthermore, we take $d = 3$ given kernel functions, namely a polynomial kernel function $k_1(a,a') = (1 + a^T a')^2$ of degree 2 for $\mathcal{K}_1$, a Gaussian kernel function $k_2(a,a') = \exp(-\frac{1}{2}(a-a')^T(a-a')/\frac{1}{10})$ for

$\mathcal{K}_2$ and a linear kernel function $k_3(a, a') = a^T a'$ for $\mathcal{K}_3$. The resulting kernel matrices are normalised according to [13, Section 4.8], giving

$$r_i = \operatorname{trace}(\mathcal{K}_i) = n + l.$$

The model parameter $c > 0$ is chosen to be

$$c = \sum_{i=1}^{d} r_i = d(n + l),$$

and we set $C = 1$.

On this application we test the three proposed versions of OGAProx. We refer to the version of OGAProx with constant parameters from Section 3.3.1 as OGAProx-C1, to the one with adaptive parameters from Section 3.3.2 as OGAProx-A and to the one from Section 4.3 giving linear convergence with constant parameters as OGAProx-C2. The results are compared with those obtained by APD1 and APD2 from [11]. In their experiments on multi kernel SVMs they showed superiority of their method compared to *Mirror Prox* by [19] in terms of accuracy, runtime and relative error. They also argued that with APD they are able to obtain decent approximations of solutions of (52) by interior point methods such as MOSEK [18] taking about the same amount of runtime.

The main difference between APD and our method OGAProx is that for the first a gradient step in the first component is employed whereas for the latter a purely proximal step is used. To be able to employ APD2 with adaptive parameters for $\nu > 0$, the roles of x and y in (54) have to be switched, giving a different method than OGAProx-A. The runtime of both methods however is still very similar as both use the same number of gradient computations/storages and projections per iteration.

All algorithms are initialised with

$$x_0 = x_{-1} = \frac{1}{d} e, \qquad y_0 = y_{-1} = 0.$$

Each data set is randomly partitioned into 80 % training and 20 % test set. The test set is used to judge the quality of the obtained model by predicting the labels via (53) and computing the resulting test set accuracy (TSA). Note that the TSA is not guaranteed to converge or increase at all by our theoretical considerations, which only state convergence of the iterates and in terms of function values. The reported TSA values are the average over 10 random partitions. Due to occasionally occurring rather dramatic deflections of the TSA we actually compute 12 runs, but remove minimum and maximum values before calculating the mean.

5.2.1 1-norm soft margin classifier

For $\mu = \nu = 0$ the formulation (52) realises the so-called 1-norm soft margin classifier. In this case g is merely convex and we can only use the constant parameter choice from Section 3.3.1 with the name OGAProx-C1. We compare the results with those obtained by APD1 from [11].

In the case of 1-norm soft margin classifier the results reported in Table 1 paint a clear picture. OGAProx outperforms APD on three out of four data sets and ties on one data set, achieving maximum TSA values of 97.45 %, 82.78 %, 93.24 % and 85.95 % on Breast cancer, Heart disease, Ionosphere and Sonar, respectively.

5.2.2 2-norm soft margin classifier

For $\mu = 0$ and $\nu > 0$ from (52) we obtain the so-called 2-norm soft margin classifier with $C = 1$. In this case g is ν-strongly convex and we can use both parameter choices from Section 3.3.1 and the one from Section 3.3.2 giving OGAProx-C1 and OGAProx-A, respectively. This time we compare the results with those obtained by APD1 as well as APD2 from [11].

We see in Table 2 that the situation for the 2-norm soft margin classifier is more diverse than previously with the 1-norm soft margin classifier. Comparing the two constant methods – OGAProx-C1 and APD1

Method	Data set	TSA at iteration k				
		$k = 250$	$k = 500$	$k = 1000$	$k = 1500$	$k = 2000$
OGAProx-C1	Breast cancer	97.15	97.37	97.08	93.94	**97.45**
	Heart disease	74.63	74.07	80.00	81.30	**82.78**
	Ionosphere	70.85	85.35	90.28	87.46	**93.24**
	Sonar	70.00	75.24	83.81	84.52	**85.95**
APD1	Breast cancer	97.23	97.37	**97.45**	94.01	**97.45**
	Heart disease	74.63	72.59	81.85	80.74	82.41
	Ionosphere	70.85	85.35	85.49	88.73	92.68
	Sonar	70.00	74.76	81.67	84.76	84.52

Table 1: TSA of 1-norm soft margin classifier ($\mu = 0$, $\nu = 0$, $C = 1$) trained with OGAProx-C1 and APD1, averaged over 10 random partitions.

Method	Data set	TSA at iteration k				
		$k = 250$	$k = 500$	$k = 1000$	$k = 1500$	$k = 2000$
OGAProx-C1	Breast cancer	97.15	97.37	97.15	**97.45**	97.15
	Heart disease	75.19	75.00	77.78	83.52	83.52
	Ionosphere	70.99	85.35	89.86	87.89	91.27
	Sonar	70.71	77.86	81.90	85.71	**86.19**
APD1	Breast cancer	97.23	97.37	97.30	97.37	97.37
	Heart disease	75.37	67.78	80.74	82.22	**84.81**
	Ionosphere	71.27	85.35	88.87	89.72	**92.39**
	Sonar	70.48	76.43	83.33	84.76	85.71
OGAProx-A	Breast cancer	97.15	97.37	97.37	97.45	97.45
	Heart disease	76.11	73.70	83.70	81.30	**84.26**
	Ionosphere	70.85	85.21	86.34	90.42	**93.52**
	Sonar	70.48	76.90	83.33	82.62	84.76
APD2	Breast cancer	97.23	97.37	**97.59**	97.01	96.72
	Heart disease	76.11	71.30	81.48	78.70	83.15
	Ionosphere	71.13	85.35	84.79	84.93	90.42
	Sonar	70.24	75.95	84.05	84.52	**86.19**

Table 2: TSA of 2-norm soft margin classifier ($\mu = 0$, $\nu = \frac{1}{2}$, $C = 1$) trained with OGAProx-C1, OGAProx-A, APD1 and APD2, averaged over 10 random partitions.

– with each other, as well as the two adaptive methods – OGAProx-A and APD2 – we see that in both cases two out of four times OGAProx is better than APD and vice versa. Notice that the two data sets with in general lower TSA, namely Heart disease and Sonar, seem to benefit from the regularising effect of $\nu > 0$, while those with already very good results on the other hand do not, compared to the results of the 1-norm soft margin classifier with $\nu = 0$. In addition note that the adaptive variant OGAProx-A improves on the result of OGAProx-C1 on three out of four data sets.

5.2.3 Regularised 2-norm soft margin classifier

For $\mu > 0$ and $\nu > 0$ from (52) we again obtain the so-called 2-norm soft margin classifier with $C = 1$, this time, however, in a regularised version. Now not only g is strongly convex, but also $\Phi(\,\cdot\,, y)$ and we can use all our parameter choices from Section 3.3.1, Section 3.3.2 and Section 4.3 yielding OGAProx-C1,

OGAProx-A and OGAProx-C2, respectively. Once more we compare the results with those obtained by APD1 as well as APD2 from [11], pointing out that that OGAProx-C2 has no APD counterpart harnessing the additional strong convexity of the problem.

Method	Data set	TSA at iteration k				
		$k = 250$	$k = 500$	$k = 1000$	$k = 1500$	$k = 2000$
OGAProx-C1	Breast cancer	97.15	97.37	97.15	**97.52**	97.45
	Heart disease	75.19	73.52	77.22	83.15	83.70
	Ionosphere	70.99	85.35	87.89	91.41	**91.97**
	Sonar	70.48	78.81	83.33	84.76	**85.95**
APD1	Breast cancer	97.23	97.37	97.37	97.01	97.30
	Heart disease	75.19	68.89	75.56	79.81	**84.07**
	Ionosphere	71.27	85.35	86.06	89.15	91.69
	Sonar	70.71	76.43	83.10	85.48	85.48
OGAProx-A	Breast cancer	97.15	97.37	97.45	97.37	97.30
	Heart disease	76.11	70.93	82.78	80.74	83.52
	Ionosphere	70.85	85.21	85.92	89.86	**93.38**
	Sonar	70.24	76.43	82.86	86.19	86.19
APD2	Breast cancer	97.23	97.37	97.45	94.53	**97.52**
	Heart disease	76.11	71.67	80.00	79.26	**83.52**
	Ionosphere	71.13	85.35	86.90	92.39	91.13
	Sonar	70.24	75.00	82.62	84.52	**86.43**
OGAProx-C2	Breast cancer	97.15	97.45	**97.59**	97.15	96.57
	Heart disease	74.07	78.52	76.11	82.22	**83.70**
	Ionosphere	70.42	84.37	86.48	90.85	**92.25**
	Sonar	69.05	74.29	85.24	85.71	**86.19**

Table 3: TSA of regularised 2-norm soft margin classifier ($\mu = 1$, $\nu = \frac{1}{2}$, $C = 1$) trained with OGAProx-C1, OGAProx-A, OGAProx-C2, APD1 and APD2, averaged over 10 random partitions.

We see in Table 3 that for the regularised 2-norm soft margin classifier the situation is similar to the version without additional regulariser. This time for the constant methods, OGAProx-C1 and APD1, OGAProx is better than APD on three data sets while APD is better than OGAProx on only one. On the contrary, for the adaptive methods, OGAProx-A and APD2, it is the other way round. APD performs better than APD on three data sets while OGAProx is better than APD on only one. For the second version of OGAProx with constant parameter choice exhibiting linear convergence in both iterates and function values, there is no APD counterpart. When we compare the results for OGAProx-C2 to those of OGAProx-C1, then we see that the TSA values become better in general with improvements on three out of four data sets and one draw. On the Breast cancer data set OGAProx-C2 even delivers the maximum TSA over all considered methods.

5.3 Classification incorporating minimax group fairness

We want to classify labelled data $(a_j, b_j)_{j=1}^n \subseteq \mathbb{R}^d \times \{\pm 1\}$, additionally taking into account so-called *minimax group fairness* [17, 7]. The data is divided into m groups $G_1, ..., G_m$, such that for $i \in [m] := \{1, ..., m\}$ we have $G_i = (a_{i_j}, b_{i_j})_{j=1}^{n_i} \subseteq (a_j, b_j)_{j=1}^n$ with $n_i := |G_i|$ and $i_j \in [n]$ for all $i \in [m]$ and all $j \in [n_i]$. *Fairness* is measured by worst-case outcomes across the considered groups. Hence we consider the following problem,

$$\min_{x \in \mathbb{R}^d} \max_{i \in [m]} f_i(x), \tag{55}$$

with

$$f_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} L(h_x(a_{i_j}), b_{i_j}),$$

where h_x is a function parametrised by x, mapping features to predicted labels, and L is a loss function measuring the error between the predicted and true labels.

It is easy to see that (55) is equivalent to

$$\min_{x \in \mathbb{R}^d} \max_{y \in \Delta_m} \sum_{i=1}^{m} y_i f_i(x),$$

where $\Delta_m := \{(v_1, ..., v_m) \in \mathbb{R}^m \mid \sum_{i=1}^m v_i = 1,\ v_i \geq 0 \text{ for } i = 1, ..., m\}$ denotes the probability simplex in $\mathbb{R}^m$. We will work with a linear (affine) predictor $h_x : \mathbb{R}^d \to \mathbb{R}$ given by

$$h_x(a) = a^T x,$$

with $x \in \mathbb{R}^d$ and $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ being the hinge loss, i.e.,

$$L(r, s) = \max\{0, 1 - sr\},$$

for r, $s \in \mathbb{R}$.

Combining all of the above we get

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^m} \Phi(x, y) - g(y), \tag{56}$$

with $\Phi : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ defined by

$$\Phi(x, y) = \sum_{i=1}^{m} y_i \frac{1}{n_i} \sum_{j=1}^{n_i} \max\{0, 1 - b_{i_j} a_{i_j}^T x\},$$

and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ given by

$$g(y) = \delta_{\Delta_m}(y).$$

The function g is proper, lower semicontinuous and convex (with modulus $\nu = 0$). Furthermore, we observe that $\Phi(\cdot, y) : \mathbb{R}^d \to \mathbb{R}$ is proper, convex and lower semicontinuous for all $y \in \operatorname{dom} g = \Delta_m$ and for all $x \in \operatorname{Pr}_{\mathbb{R}^d}(\operatorname{dom} \Phi) = \mathbb{R}^d$ we have $\operatorname{dom} \Phi(x, \cdot) = \mathbb{R}^m$ and $\Phi(x, \cdot) : \mathbb{R}^m \to \mathbb{R}$ is concave and Fréchet differentiable. However, note that Φ is not differentiable in its first component.

Moreover the Lipschitz condition on the gradient is fulfilled as well. Indeed, for (x, y), $(x', y') \in \mathbb{R}^d \times \Delta_m$ we have

$$\left\| \nabla_y \Phi(x, y) - \nabla_y \Phi(x', y') \right\| \leq L_{yx} \left\| x - x' \right\| + L_{yy} \left\| y - y' \right\|,$$

with

$$L_{yx} = \sqrt{\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} \left\| a_{i_j} \right\|^2} \quad \text{and} \quad L_{yy} = 0.$$

Additionally, with $\tau > 0$ and $y \in \operatorname{dom} g$, we have for $x \in \mathbb{R}^d$

$$\operatorname{prox}_{\tau \Phi(\cdot, y)}(x) = \arg\min_{u \in \mathbb{R}^d} \left\{ \tau \sum_{i=1}^{m} y_i \frac{1}{n_i} \sum_{j=1}^{n_i} \max\{0, 1 - b_{i_j} a_{i_j}^T u\} + \frac{1}{2} \|u - x\|^2 \right\}.$$

By introducing slack variables for the pointwise maximum, we see that the above minimisation problem is equivalent to the following quadratic program

$$\min_{\substack{u \in \mathbb{R}^d, \\ r_{ij} \in \mathbb{R}, \\ i \in [m],\ j \in [n_i]}} \left\{ \tau \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_i \frac{1}{n_i} r_{ij} + \frac{1}{2} \|u - x\|^2 \right\}.$$

$$\text{s.t.} \quad r_{ij} \geq 0 \quad \forall\, i \in [m],\ \forall\, j \in [n_i]$$

$$r_{ij} + b_{i_j} a_{i_j}^T u \geq 1 \quad \forall\, i \in [m],\ \forall\, j \in [n_i]$$

	Group S1		Group S2		Overall	
k	with fairness	without fairness	with fairness	without fairness	with fairness	without fairness
100	**95.78**	**95.78**	80.68	80.84	85.56	85.56
500	**95.78**	**95.78**	**81.15**	80.28	**85.93**	85.19
1000	**95.78**	**95.78**	**81.15**	80.28	**85.93**	85.19

Table 4: TSA of the affine classifier after k iterations of OGAProx for the groups according to "sex", averaged over 5 random partitions.

	Group A1		Group A2		Group A3		Overall	
k	with fairness	without fairness	with fairness	without fairness	with fairness	without fairness	with fairness	without fairness
100	87.76	86.48	82.97	82.97	**86.93**	**86.93**	85.93	85.56
500	**88.71**	85.53	**83.84**	82.97	**86.93**	**86.93**	**86.67**	85.19
1000	**88.71**	85.53	**83.84**	82.97	**86.93**	**86.93**	**86.67**	85.19

Table 5: TSA of the affine classifier after k iterations of OGAProx for the groups according to "age", averaged over 5 random partitions.

For our practical applications we consider the Statlog *heart disease* data set (270 observations; 13 features) from the "UCI Machine Learning Repository" [8] and consider two different groupings; one consisting of the sex of the patients, while the other one is regarding the patients' age. For "sex" we have two groups, that is female patients (Group S1) and male patients (Group S2), whereas for "age" we consider three groups, that is patients that are younger than 50 years old (Group A1), patients that are younger than 60 but at least 50 years old (Group A2), and patients that are 60 years of age or older (Group A3). The data set is randomly partitioned into 80 % training data and 20 % test data. The results in Table 4 and Table 5 are the values of the achieved test set accuracy (TSA) averaged over 5 random partitions. For each considered group we state the intragroup TSA together with the overall TSA for the entire test set.

Every time we report the results obtained by iterates of OGAProx governed by solving the minimax problem (56) taking into account the considered groups ("with fairness"), as well as the results obtained by not taking into account minimax group fairness ("without fairness"), i.e., solving the problem for a single extensive group $G_1 = (a_j, b_j)_{j=1}^n$ with $n_1 = n$, yielding the minimisation of the average loss over the whole population and leading to an "ordinary" minimisation problem.

We see in Table 4 and Table 5 that taking into account the groups regarding "sex" and "age", respectively, is beneficial for training the affine classifier. In both cases "with fairness" achieves the highest TSA for each group and at the same time the highest overall TSA as well.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Acknowledgments

The work of RIB and of ERC is supported by FWF (Austrian Science Fund), projects W 1260 and P 29809-N32, respectively.

References

[1] Bauschke, H.H., Combettes, P.L.: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer New York, 2011.

[2] Boţ, R.I., Böhm, A: *Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems*. arXiv:2007.13605, 2020.

[3] Böhm, A., Sedlmayer, M., Csetnek, E.R., Boţ, R.I.: *Two steps at a time–taking GAN training in stride with Tseng's method*. arXiv:2006.09033, 2020.

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

[5] Daskalakis, C., Ilyas, A., Syrgkanis, V., Zeng, H.: *Training GANs with optimism*. In: International Conference on Learning Representations, 2018. https://openreview.net/forum?id=SJJySbbAZ.

[6] Daskalakis, C., Panageas, I.: *The limit points of (optimistic) gradient descent in min-max optimization*. In: Advances in Neural Information Processing Systems, 9236–9246, 2018.

[7] Diana, E., Gill, W., Kearns, M., Kenthapadi, K., Roth, A.: *Minimax group fairness: algorithms and experiments*. arXiv:2011.03108, 2020.

[8] Dua, D., Graff, C.: *UCI Machine Learning Repository*. School of Information and Computer Science, University of California, 2019. http://archive.ics.uci.edu/ml.

[9] Gidel, G., Berard, H., Vignoud, G., Vincent, P., Lacoste-Julien, S.: *A variational inequality perspective on generative adversarial networks*. In: International Conference on Learning Representations, 2019. https://openreview.net/forum?id=r1laEnA5Ym.

[10] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair S., Courville, A., Bengio, Y.: *Generative adversarial nets*. In: Advances in Neural Information Processing Systems, 2672–2680, 2014.

[11] Hamedani, E.Y., Aybat, N.S.: *A primal-dual algorithm with line search for general convex-concave saddle point problems*. SIAM Journal on Optimization 31(2):1299–1329, 2021.

[12] Korpelevich, G.M.: *The extragradient method for finding saddle points and other problems*. Ekonomika i Matematicheskie Metody 12(4):747-756, 1976.

[13] Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I.: *Learning the kernel matrix with semidefinite programming*. Journal of Machine Learning Research 5:27–72, 2004.

[14] Liang, T., Stokes, J.: *Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks*. In: K., Chaudhuri, M., Sugiyama, The 22nd International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research 89:907–915, 2019.

[15] Malitsky, Y., Tam, M.K.: *A forward-backward splitting method for monotone inclusions without cocoercivity*. SIAM Journal on Optimization 30(2):1451–1472, 2020.

[16] Mangasarian, O.L., Wolberg, W.H.: *Cancer diagnosis via linear programming*, SIAM News 23(5): 1–18, 1990.

[17] Martinez, N., Bertran, M., Sapiro, G.: *Minimax pareto fairness: A multi objective perspective*. In: International Conference on Machine Learning, 6755–6764, 2020.

[18] MOSEK ApS: *The MOSEK Optimization Toolbox for MATLAB Manual. Version 9.0*, 2019. http://docs.mosek.com/9.0/toolbox/index.html.

[19] Nemirovski, A.: *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 Journal on Optimization 15(1):229–251, 2004.

[20] Opial, Z.: *Weak convergence of the sequence of successive approximations for nonexpansive mappings*. Bulletin of the American Mathematical Society, 73(4):591–597, 1967.

[21] Rockafellar, R.T.: *Monotone operators associated with saddle-functions and minimax problems*. In: F.E., Browder (ed.), Nonlinear Functional Analysis, Proceedings of Symposia in Pure Mathematics 18:241–250, 1970.

[22] Tseng, P.: *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*. SIAM Journal on Control and Optimization, 29(1):119–138, 1991.

[23] Von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[24] Zhang, G., Wang, Y., Lessard, L., Grosse, R.: *Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization*. arXiv:2102.09468, 2021.