On the Asymptotic Convergence of Full LOLA

Raphael Chinchilla

January 2021

1 Introduction

Solving a minmax problem, also known as robust optimization, consists on finding the optimal strategies for two players that want to optimize opposite interests. Conceptually, this is a versatile paradigm that can be used to model a variety of situations, including games such chess, elections between two candidates, an airplane flying in the middle of a storm and neural network accurately classifying misleading information. Most modern algorithms to solve minmax problems have players choosing locally their strategy without taking into account what the other player will do. Our goal in this technical note is to develop an algorithm in which each player takes chooses their local strategy while taking into account what will be other player's action.

The modern approach to minmax problems was established in the seminal technical note by von Neumann [1] in which he proved that if the problem is convex in the minimization and concave in the maximization, then the min and the max commute, meaning that the order of the players does not matter. This is known as the Minmax Theorem and has since been extended to other cases [2, 3].

However, in many problems of interest, the min and max do not necessarily commute. Some of these include adversarial learning [4, 5, 6, 7], generative adversarial networks [8], robust model predictive control [9, 10, 11], robust estimation [12, 13], robust optimization for stochastic optimization [14, 15, 16], among many.

In some cases, it is possible to solve the non-commuting minmax problems using approaches such as robust counterpart or cutting-set methods [17, 18, 19, 20]. In the other cases, generally when the problem is non-convex non-concave, one is usually restricted to finding local minmax points, as defined in [21], which have first and second order necessary and sufficient conditions obtained from the gradient and Hessian.

An elegant method to look for points satisfying the first order necessary condition is the Learning with Opponent Learning Awareness (LOLA) introduced in [22]. The idea of LOLA is that the minimizer chooses its direction based on the predicted direction the maximizer will take. The convergence of a modified version of LOLA was given by the same group of authors [23].

In this technical note, we introduce the Full LOLA algorithm, of which the standard LOLA can be seen as a linearization of. In our opinion, the Full LOLA approach has several elegant properties which motivated us to explore using it. While we did not found any numerical application that could benefit from this approach instead of using Gradient Descent Ascent or other (strictly) first order methods, we believe the intuitions developed in these proofs might end up being useful either for other proofs or in applications we were not aware of.

Statement of Contributions: The main ingredient of our approach is what we call the full descent ascent directions, defined in Section 2. In essence, in a full descent ascent direction, the minimizer does not decrease the cost function at the current point, but instead decreases the cost function calculated at the next value that the maximizer will take. This choice of directions reflects the asymmetry of minmax games, in which the minimizer has less freedom to chose their action than the maximizer has. Building from this definition, in Section 3 we propose a method to obtain full descent ascent directions based on gradients. For the maximizer, this is equivalent to gradient ascent. For the minimizer, the descent direction is obtained from a modified version of the cost function, in which the value of the maximizer is offset by the gradient ascent step. The method we propose is actually slightly more general, and allow us to solve problems with convex constraints. It also allow us to use scaling matrices to compute the directions, such as the Hessian, obtaining Newton types algorithms. In Section 4 we prove the asymptotic convergence of the method, for two types of step sizes, either fixed or adjusted using an Armijo rule. For convenience of the exposition, all the proof are in the Appendix.

Notation: The set of real numbers is denoted by $\mathbb{R}$. Given a vector $v \in \mathbb{R}^n$, its transpose is denoted by v'. Consider a differentiable function $f : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}^p$. The Jacobian (or gradient if $p = 1$) at a point $(\bar{x}, \bar{y})$ according to the x variable is a matrix of size $n \times p$ and is denoted by $\boldsymbol{\nabla}_x f(\bar{x}, \bar{y})$. The partial derivative according to the coordinate x is a matrix of size $n \times p$ and is denoted by $\partial_x f(\bar{x}, \bar{y})$. Given a differentiable function $g : \mathbb{R}^n \mapsto \mathbb{R}^m$, $\boldsymbol{\nabla}_x f(\bar{x}, g(\bar{x})) = \partial_x f(\bar{x}, g(\bar{x})) + \boldsymbol{\nabla}_x g(\bar{x}) \partial_y f(\bar{x}, g(\bar{x}))$. For a twice differentiable function, the cross derivative is given by $\boldsymbol{\nabla}_{xy} f(\bar{x}, \bar{y}) = \boldsymbol{\nabla}_x(\boldsymbol{\nabla}_y f(\bar{x}, \bar{y}))$.

2 Problem statement

Consider two non-empty, closed and convex sets [1] $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ and a function $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$. The minmax optimization

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) \tag{1}$$

denotes the problem of finding a point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ such that $\forall y \in \mathcal{Y}$ and $\forall x \in \mathcal{X}$

$$f(x^*, y) \leqslant f(x^*, y^*) \leqslant \max_{\tilde{y} \in \mathcal{Y}} f(x, \tilde{y}).$$

If such point exists, it is called a global minmax of $f(\cdot)$.

2.1 Local minmax

Except in some specific cases, such as when $f(\cdot)$ is convex in x and concave in y, finding a global minmax is extremely challenging. An alternative is to look for a local minmax, which was first defined in [21].

Definition 1 (Local minmax according to Jin et al.) *A point (x^*, y^*) is said to be a local minmax of $f(\cdot)$ if there exist $\delta_0 > 0$ and a positive function $h(\cdot)$ satisfying $h(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta \in (0, \delta_0]$, $\forall x \in \mathcal{X} : \|x - x^*\| \leqslant \delta$ and $\forall y \in \mathcal{Y} : \|y - y^*\| \leqslant h(\delta)$ we have that*

$$f(x^*, y) \leqslant f(x^*, y^*) \leqslant \max_{\tilde{y} \in \mathcal{Y} : \|\tilde{y} - y^*\| \leqslant h(\delta)} f(x, \tilde{y}) \qquad \square$$

Essentially, a local minmax is defined by properties that hold on neighborhoods around (x^*, y^*). Local properties have the advantage that they tend to be easier to verify than global ones. Unfortunately, a global minmax might not be a local minmax, and we refer the reader to the original paper for a counter example and an analysis on this question.

Despite this evident drawback in the definition of local minmax, one of its main advantages is that one can deduce first order necessary conditions of optimality. We state the result in a slightly more general form than it is stated in [21] in order to take into account constraints.

Proposition 1 (First order necessary condition) *Assume $f(\cdot)$ is continuously differentiable and (x^*, y^*) is a local minmax. Then, $\forall y \in \mathcal{Y}$, $(y - y^*)' \boldsymbol{\nabla}_y f(x^*, y^*) \leqslant 0$. Moreover, $\exists \delta_0 > 0$ such that $\forall x \in \mathcal{X} : \|x - x^*\| < \delta_0$, $(x - x^*)' \boldsymbol{\nabla}_x f(x^*, y^*) \geqslant 0$.* □

Corollary 1 (Unconstrained conditions) *Assume $f(\cdot)$ is continuously differentiable and (x^*, y^*) is a local minmax and an interior point of $\mathcal{X} \times \mathcal{Y}$. Then $\boldsymbol{\nabla}_y f(x^*, y^*) = 0$ and $\boldsymbol{\nabla}_x f(x^*, y^*) = 0$.* □

[1] We remind the reader that $\mathbb{R}^n$ is closed and convex.

Proof. For the max, if y^* is an interior point of $\mathcal{Y}$, then for any $y \in \mathcal{Y}$ there is a $\beta \in (0,1]$ such that $y^* - \beta(y - y^*) \in \mathcal{Y}$. Therefore we have that $(y - y^*)' \nabla_y f(x^*, y^*) \leqslant 0$ and that $-\beta(y - y^*)' \nabla_y f(x^*, y^*) \leqslant 0$ which implies that $\nabla_y f(x^*, y^*) = 0$. The proof for the min is equivalent. ■

2.2 Descent ascent algorithms

Consider two arbitrary functions $d_x(x,y)$ and $d_y(x,y)$ that satisfy the conditions that $x + d_x(x,y) \in \mathcal{X}$ and $y + d_y(x,y) \in \mathcal{Y}$ and a sequence of scalars $\{(\alpha^k, \beta^k)\}$ with $\alpha^k, \beta^k \in (0,1]$. Given an initial point $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$, the sequence $\{(x^k, y^k)\}$ is recursively defined by

$$\begin{aligned} x^{k+1} &= x^k + \alpha^k d_x(x^k, y^k) \\ y^{k+1} &= y^k + \beta^k d_y(x^{k+1}, y^k). \end{aligned} \tag{2}$$

In general, there are no closed form expressions to obtain local minmax points. Instead, one uses descent ascent algorithms, in which one designs numerical functions $d_x(x,y)$ and $d_y(x,y)$ and sequences $\{(\alpha^k, \beta^k)\}$ such that every limit point of the sequence $\{(x^k, y^k)\}$ satisfies the first order optimality conditions of Proposition 1; **we call such points of stationary points**. The most common type of descent ascent algorithms uses alternating descent ascent sequences, for which we give the following definition:

Definition 2 (Alternating descent ascent) *We say that the sequence defined by* (2) *is an alternating descent ascent sequence if it satisfies*

$$f(x^{k+1}, y^k) \leqslant f(x^k, y^k) \tag{3a}$$

$$f(x^{k+1}, y^{k+1}) \geqslant f(x^{k+1}, y^k). \tag{3b}$$

with at least one of the inequalities holding strictly. By extension, we say that $\alpha d_x(x,y)$ and $\beta d_y(x,y)$ are alternating descent ascent directions. □

This formality includes many of the most popular algorithms minmax algorithms. Here are some examples:

1. Gradient Descent Ascent: $d_x(x,y) = -\alpha \nabla_x f(x,y)$ and $d_y(x,y) = \beta \nabla_y f(x,y)$ with α and $\beta \in (0, +\infty)$
2. Gradient Descent multiple Ascent: $d_x(x,y) = -\alpha \nabla_x f(x,y)$ and $d(x,y) = \sum_{k=1}^{n} \nabla_y f(x, \tilde{y}_k)$, with α and $\beta \in (0, +\infty)$ and where $\nabla_y f(x, \tilde{y}_k)$ is implicitly defined by

$$\begin{aligned} \tilde{y}_1 &= y + \beta \nabla_y f(x,y) \\ \tilde{y}_2 &= \tilde{y}_1 + \beta \nabla_y f(x, \tilde{y}_1) \\ &\vdots \end{aligned}$$

3. GradaMax: $d_x(x,y) = -\alpha \nabla_x f(x,y)$ with $\alpha \in (0, +\infty)$ and $d_y(x,y) \in \arg\max_{d_y : y + d_y \in \mathcal{Y}} f(x, y + d_y)$, $\beta = 1$.
4. Alternating minmax: $d_x(x,y) \in \arg\min_{d_x : x + d_x \in \mathcal{X}} f(x + d_x, y)$ and $d_y(x,y) \in \arg\max_{d_y : y + d_y \in \mathcal{Y}} f(x, y + d_y)$.

Other methods popular in the robust training community such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) can also be expressed as alternating directions minmax.

A notable characteristic of alternating descent ascent sequences is that each player takes an action without taking into consideration what will be the consequences on the other player's action. Instead, we argue in this technical note for an approach where $d_x(x,y)$ and $d_y(x,y)$ are computed simultaneously, each player choosing their action while taking into account the other player's move. This is captured in the following definition.

Definition 3 (Full descent ascent) *We say that the sequence defined by* (2) *is a full descent ascent sequence if it satisfies*

$$f(x^{k+1}, y^{k+1}) \leqslant f(x^k, y^k + \beta^k d_y(x^k, y^k)) \tag{4a}$$
$$f(x^{k+1}, y^{k+1}) \geqslant f(x^{k+1}, y^k). \tag{4b}$$

with at least one of the inequalities holding strictly. By extension, we say that $\alpha d_x(x,y)$ and $\beta d_y(x,y)$ are full descent ascent directions. □

Fundamentally, the full descent ascent captures the nature of minmax optimizations. Not only the descent ascent step choices are, by construction, asymmetric, but it also reflects the fact the minimization needs to chose their step considering what will be the action of the max.

Remark 1 (Solving minmax as a full descent ascent algorithm) *If one uses the GradMax (as defined above), then the sequence* (2) *could asymptotically converge towards a local minmax, most notable if $f(\cdot)$ is strongly convex in x and strongly concave in y. Now, consider an analogous choice of full descent ascent directions given by:*

$$d_x(x^k, y^k) \in \underset{d_x : d_x + x^k \in \mathcal{X}}{\arg\min} \; f(x + d_x, y + d_y(x + d_x, y))$$
$$d_y(x^{k+1}, y^k) = \underset{d_y : y + d_y \in \mathcal{Y}}{\arg\max} \; f(x^{k+1}, y^k + d_y).$$

where we assume the arg max *is uniquely achieved. This choice of directions is* ***exactly*** *the solution of the minmax optimization. Evidently, one does not have access to closed form expressions of such functions, as this is the goal itself of an optimization algorithm. However, this shows how the full descent ascent directions describe a more appropriate concept of direction to find mimnax points.*

3 Obtaining local full descent ascent directions

In order to obtain local $d_x(x,y)$ and $d_y(x,y)$, it is usefull to consider the following result from minimization. Suppose one wants to solve the problem $\min_{x \in \mathcal{X}} f(x)$ where $\mathcal{X}$ is a convex set. If $f(\cdot)$ is continuously differentiable, projected direction methods solve this optimization by generating a sequence $x^{k+1} = x^k + \alpha d_x(x^k)$ where $d_x(x)$ is a local descent direction obtained from solving the quadratic subproblem

$$d_x(x) = \underset{d_x : d_x + x \in \mathcal{X}}{\arg\min} \; f(x) + d_x' \boldsymbol{\nabla}_x f(x) + \frac{1}{2} d_x' A(x) d_x \tag{5}$$

where $A(x)$ is a strictly positive definite matrix and $\alpha \in (0, 1]$. A large number of optimization methods can be written in this form including gradient descent (choosing $A(x)$ as the identity matrix), Newton method (choosing $A(x)$ as the Hessian matrix), Gauss-Newton method and its generalizations, Quasi-Newton methods, Trust Region methods (by also including a constraint on the norm of d_x) among many others.

In an analogous way, if $f(x,y)$ is differentiable in y, we define $d_y(x,y)$ as the solution of

$$d_y(x,y) = \underset{d_y : y + d_y \in \mathcal{Y}}{\arg\max} \, f(x,y) + d_y' \boldsymbol{\nabla}_y f(x,y) - \frac{1}{2} d_y' B(x,y) d_y \tag{6}$$

where $B(x,y)$ is a positive definite matrix. It is important to emphasize that $d_y(x,y)$ is function both of x and y. Consider the function $\hat{f}_x(x,y)$ defined by

$$\hat{f}_x(x,y) := f(x, y + \beta d_y(x,y)). \tag{7}$$

If $\hat{f}_x(x,y)$ is differentiable in x, we define $d_x(x,y)$ by

$$d_x(x,y) = \underset{d_x : x + d_x \in \mathcal{X}}{\arg\min} \hat{f}_x(x,y) + d_x' \boldsymbol{\nabla}_x \hat{f}_x(x,y) + \frac{1}{2} d_x' A(x,y) d_x \tag{8}$$

where $A(x,y)$ is a positive definite matrix. We can now state our first result

Proposition 2 (Computing local directions) *If $f(x,y)$ is continuously differentiable with respect to y and $\hat{f}_x(x,y)$ is continuously differentiable with respect to x on a neighborhood around a point $(\tilde{x},\tilde{y})$ which is not a stationary point, then there exist α_0 and β_0 such that $\forall\alpha \in (0,\alpha_0)$ and $\forall\beta \in (0,\beta_0)$ $\alpha d_x(\tilde{x},\tilde{y})$ and $\beta d_y(\tilde{x},\tilde{y})$ are full descent ascent directions.* □

We will now look at two particular choices of matrices $A(x,y)$ and $B(x,y)$ that will also help understanding the algorithm. In both we will consider the unconstrained case ($\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$).

3.1 Full LOLA

The first case is when one chooses $A(x,y)$ and $B(x,y)$ as the identity matrix, which is what we call the full LOLA. The direction for the max is

$$\beta d_y(x,y) = \beta\boldsymbol{\nabla}_y f(x,y).$$

For the min, the direction is

$$\begin{aligned} d_x(x,y) =& -\boldsymbol{\nabla}_x f(x, y + \beta\boldsymbol{\nabla}_y f(x,y)) \\ =& -\partial_x f(x, y + \beta\boldsymbol{\nabla}_y f(x,y)) - \beta\boldsymbol{\nabla}_{xy} f(x,y)\partial_y f(x, y + \beta\boldsymbol{\nabla}_y f(x,y)). \end{aligned}$$

If one linearizes this direction around (x,y) one obtains

$$d_x(x,y) = -\boldsymbol{\nabla}_x f(x,y) - \beta\boldsymbol{\nabla}_{xy} f(x,y)\boldsymbol{\nabla}_y f(x,y)$$

i.e., the standard LOLA direction.

Using these results, the full descent ascent sequence is

$$\begin{aligned} x^{k+1} &= x^k + \alpha^k - \boldsymbol{\nabla}_x f(x, y + \beta\boldsymbol{\nabla}_y f(x,y)) \\ y^{k+1} &= y^k + \beta^k\boldsymbol{\nabla}_y f(x^{k+1}, y^k). \end{aligned}$$

In contrast with the standard (alternating) gradient descent ascent, in (full) LOLA, the descent direction uses the gradient of the maximzer to correct the direction towards where it should go. In the case where case where (x^k, y^k) is a local maximum, both the full and standard LOLA are equivalent to a gradient descent ascent as $\boldsymbol{\nabla}_y f(x^k, y^k) = 0$.

3.2 Full Newton types algorithms

Full descent ascent algorithms can also be used as Newton types algorithms by choosing matrices $A(x,y)$ and $B(x,y)$ as Hessian. For the maximizer, the straightforward choice of matrix is $B(x,y) = -\boldsymbol{\nabla}_{yy} f(x,y)$. For the minimizer there are two options. The first option is to take $A(x,y) = \boldsymbol{\nabla}_{xx} f(x,y)$ and the secondis to take

$$A(x,y) = \boldsymbol{\nabla}_{xx} f\Big(x, y - \beta\boldsymbol{\nabla}_{yy} f(x,y)^{-1}\boldsymbol{\nabla}_y f(x,y)\Big)$$

Taking $A(x,y) = \boldsymbol{\nabla}_{xx} f(x,y)$ has the advantage of making the differentiation easier, while in the second option we more closely maintain the spirit of full descent ascent of minimizing the cost of the future direction.

Remark 2 (Differentiability of $\hat{f}_x(x,y)$) *The assumption of differentiability of $\hat{f}_x(x,y)$ with respect to x is closely related to the differentiability of $d_y(x,y)$, which is known as sensitivity analysis. In the case where (x,y) is an interior point of the constrain set (or, equivalently, if $\mathcal{Y} = \mathbb{R}^m$) a sufficient condition is that $\boldsymbol{\nabla}_y f(x,y)$ and $B(x,y)$ are differentiable. However, establishing differentiability in the case where (x,y) is not an interior point is substantially more challenging, and naming such conditions goes beyond the scope of this technical note. We refer the reader to [24] which has a thorough treatment of the topic.* □

Remark 3 (Using momentum) *In minimization, algorithms with momentum are of the general form $x^{k+1} = x^k - \alpha(\boldsymbol{\nabla}_x f(x^k) + p_x^k)$. One example of such algorithm is to use $p_x^k = \boldsymbol{\nabla}_x f(x^{k-1}) + \mu p_x^{k-1}$ with $\mu \in [0,1]$.*

The framework of full descent ascent also allows for methods with momentum by substituting $\boldsymbol{\nabla}_y f_x(x,y)$ by $\boldsymbol{\nabla}_y f_x(x,y) + p_y$ in (6) and $\boldsymbol{\nabla}_x \hat{f}_x(x,y)$ by $\boldsymbol{\nabla}_x \hat{f}_x(x,y) + p_x$ in (8), although these might no longer be full descent ascent directions as we define in Definition 3. □

4 Asymptotic convergence

Our goal now is to obtain conditions such that every limit point of the sequence

$$\begin{aligned} x^{k+1} &= x^k + \alpha^k d_x(x^k, y^k) \\ y^{k+1} &= y^k + \beta^k d_y(x^{k+1}, y^k) \end{aligned}$$

where $d_x(x^k, y^k)$ is given by (8) and $d_y(x^{k+1}, y^k)$ is given by (6) is a stationary point. We will not make any assumption of convexity or concavity instead casting the results in the most general possible way. For this reason, our convergence results will pertain to asymptotic properties of full descent ascent sequences. Results on non asymptotic convergence will be the subject of a future work.

For the sake of conciseness, we will state our using the notation

$$\hat{f}_x(x, y) = f(x, y + \beta d_y(x, y)).$$

Let us denote by $\lambda_{min}(M)$ and $\lambda_{max}(M)$ the smallest and largest eigenvalues of a symmetric matrix M. In addition to the assumptions of convexity and closeness of $\mathcal{X}$ and $\mathcal{Y}$ and the continuous differentiability of $f(\cdot)$ and $\hat{f}_x(\cdot)$ we will also need the following assumptions.

Assumption 1 *Given a full descent ascent sequence $\{(x^k, y^k)\}$, for all k the eigenvalues of $A(x^k, y^k)$ and $B(x^k, y^k)$ are bounded by bellow and above and away from zero, meaning that there exist positive constants c_1, c_2, c_3, c_4 such that $\forall k > 0$,*

$$\begin{aligned} &\lambda_{min}(A(x^k, y^k)) > c_1 & &\lambda_{max}(A(x^k, y^k)) < c_2 \\ &\lambda_{min}(B(x^k, y^k)) > c_3 & &\lambda_{max}(B(x^k, y^k)) < c_4 \end{aligned} \qquad \square$$

This assumption essentially guarantees that optimizations (6) and (8) will always be well defined and only have one solution. It is important to emphasize that $A(x, y)$ and $B(x, y)$ are algorithmic choices in the sense that they are chosen by the practitioner.

Our first result concerns the convergence when the matrices $A(x^k, x^k)$ and $B(x^k, x^k)$ and the step sizes α^k, β^k are constant.

Theorem 1 (Constant step size) *Let $\{(x^k, y^k)\}$ be a full descent ascent sequence with $\alpha^k = \beta^k = 1$, $A(x, y) = A$ and $B(x, y) = B$. Assume that $\forall x_1, x_2 \in \mathcal{X}$ and $\forall y_1, y_2 \in \mathcal{Y}$ there exist constants $L_x, L_y > 0$ such that the following smoothness condition holds:*

$$\|\boldsymbol{\nabla}_x f(x_1, y_1) - \boldsymbol{\nabla}_x f(x_2, y_2)\| < L_x \sqrt{\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2}$$

$$\|\boldsymbol{\nabla}_y f(x_1, y_1) - \boldsymbol{\nabla}_y f(x_2, y_2)\| < L_y \sqrt{\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2}$$

If $2\left(L_x \sqrt{1 + \lambda_{min}(B)\, L_y^2}\right)^{-1} > \lambda_{min}(A)^{-1}$ and $2\, L_y^{-1} > \lambda_{min}(B)^{-1}$ then every limit point of $\{(x^k, y^k)\}$ is a stationary point of $f(\cdot)$. Moreover, $d_x(x, y)$ and $d_y(x, y)$ are full descent ascent directions, meaning that

$$\begin{aligned} f(x^{k+1}, y^{k+1}) &\leqslant f(x^k, y^k + d_y(x^k, y^k)) \\ f(x^{k+1}, y^{k+1}) &\geqslant f(x^{k+1}, y^k) \end{aligned}$$

with at least one of the inequalities holding strictly. □

It is easier to interpret Theorem 1 when $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}^m$ The full descent ascent directions are

$$\begin{aligned} d_y(x, y) =& B^{-1} \boldsymbol{\nabla}_y f(x, y) \\ d_x(x, y) =& - A^{-1} \boldsymbol{\nabla}_x f(x, y + B^{-1} \boldsymbol{\nabla}_y f(x, y)). \end{aligned}$$

Algorithm 1 Simultaneous descent ascent with Armijo rule

```
Require: An initial point $(x^0, y^0)$ and rates $r_x, r_y \in (0,1)$
 1: $(\alpha^k, \beta^k) = (1,1)$
 2: while $f(\tilde{x},\tilde{y}) - F_{min} > 0$ and $f(\tilde{x},\tilde{y}) - F_{max} < 0$ do
 3:     $\tilde{x} = x^k + \alpha^k d_x(x^k, y^k)$
 4:     $\tilde{y} = y^k + \beta^k d_y(\tilde{x}, y^k)$
 5:     $F_{min} = \hat{f}_x(x^k, y^k) + \sigma_x \alpha^k d_x(x^k, y^k)' \boldsymbol{\nabla}_x \hat{f}(x^k, y^k)$
 6:     $F_{max} = f(\tilde{x}, y^k) + \sigma_y \, \beta^k \, d_y(\tilde{x}, y^k)' \boldsymbol{\nabla}_y f(\tilde{x}, y^k)$
 7:     if $f(\tilde{x},\tilde{y}) - F_{min} > 0$ then
 8:         $\alpha^k = \alpha^k \, r_x$
 9:     end if
10:     if $f(\tilde{x},\tilde{y}) - F_{max} < 0$ then
11:         $\beta^k = \beta^k \, r_y$
12:     end if
13: end while
14: $x^{k+1} = \tilde{x}$
15: $y^{k+1} = \tilde{y}$
16: $k = k + 1$
17: Go to 1
```

Now if we use the fact that $\lambda_{max}(A^{-1}) = \lambda_{min}(A)^{-1}$ and equivalent to B, Theorem 1 essentially says two things. The first one is that the larger the constants L_x, L_y are, the smaller the step sizes, represented by $\lambda_{min}(A)^{-1}$ and $\lambda_{min}(B)^{-1}$, can be. This kind of result is typical in optimization. But the second particularly interesting thing is that the maximum step size of the minimizer depends on the step size of maximizer, essentially stating that if the maximizer take small steps, the minimizer also needs to take small steps. The idea that the minimizer needs to take smaller steps than the maximizer is common in minmax optimization (see for instance the discussion for Gradient Descent Ascent on [21]). What is innovative in our result is that we are able to quantify exactly how big the step can be.

The biggest limitation of Theorem 1 is that one often does not know the values of L_x and L_y. As a consequence, one would need to manually tune the matrices A and B using a on trial and error, and the step sizes are rarely as large as they could be. Our next result uses an Armijo type condition and a backtracking algorithm to determine the step sizes. We point out that the result does not require the smoothness condition.

Take two scalars $\sigma_x, \sigma_y \in (0,1)$. At a given point (x^k, y^k), given two step sizes (α^k, β^k), we define the following Armijo type conditions for the minmax

$$f(x^{k+1}, y^{k+1}) - \hat{f}_x(x^k, y^k) \leqslant \sigma_x \, \alpha^k \, d_x(x^k, y^k)' \boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k) \tag{9a}$$

$$f(x^{k+1}, y^{k+1}) - f(x^{k+1}, y^k) \geqslant \sigma_y \, \beta^k \, d_y(x^{k+1}, y^k)' \boldsymbol{\nabla}_y f(x^k, y^k) \tag{9b}$$

with at least one of the inequalities holding strictly and where $\hat{f}_x(x, y)$ is given by (7). We bring attention to the reader that $d_x(x, y)$ and $\hat{f}(x, y)$ depend on the value of β^k. These Armijo conditions not only guarantee that $\alpha^k d_x(x^k, y^k)$ and $\beta^k d_y(x^{k+1}, y^k)$ are full descent ascent directions, but also guarantees that at each iteration the steps are sufficiently large. We use these conditions to design Algorithm 1 and prove its convergence.

Theorem 2 (Convergence of Armijo) *Every limit point of a sequence $\{(x^k, y^k)\}$ generated by Algorithm 1 is a stationary point.* □

The idea behind Algorithm 1 is to obtain steps sizes α^k, β^k that satisfy the Armijo conditions by implementing a backtracking algorithm. A fundamental aspect of the algorithm is that α^k and β^k are updated only when they do not satisfy their respective Armijo conditions; this plays a crucial role in the proof of Theorem 2.

Theorem 1 and Theorem 2 guarantee that every limit point of the generated full descent ascent sequence is a stationary point, but they do not guarantee that such limit points exist. This is guaranteed by the next result, the Capture Theorem. The Capture Theorem essentially says that, if $(x^* y^*)$ is an isolated local minmax, if one element $(x^{\bar{k}}, y^{\bar{k}})$ of the full descent ascent passes close enough to it, then $\{(x^k, y^k)\}$ will converge towards (x^*, y^*).

Theorem 3 (Capture Theorem) *Let $\{(x^k, y^k)\}$ be a sequence generated by the full descent ascent direction method using either the Theorem 1 or Theorem 2. Let (x^*, y^*) be an isolated local minmax on a neighborhood where it is also the only stationary point. Then there exist a neighborhood $S_x \subset \mathcal{X}$ around x^* and a neighborhood $S_y \subset \mathcal{Y}$ around y^* such that if for some $\bar{k}$, $(x^{\bar{k}}, y^{\bar{k}}) \in S_x \times S_y$ then $\lim_{k,k>\bar{k}}(x^k, y^k) = (x^*, y^*)$.* □

5 Conclusion

In this technical note, we have presented a new type of algorithm to solve minmax optimization using what we call full descent ascent directions. We have shown that such directions are better at generalizing the concept of descent direction from regular optimization. We were also able to state conditions that guarantee the asymptotic convergence of such algorithm to local minmax points.

While we have not found applications for which full descent ascent directions outperform the state of the art, they provide an elegant way to look at minmax optimization. Further exploration, both with respect to the theory and practice, could unfold cases in which such directions outperform other methods.

References

[1] J. von Neumann, "Zur theorie der gesellschaftsspiele," *Mathematische annalen*, vol. 100, no. 1, pp. 295–320, 1928.

[2] K. Fan, "Minimax theorems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 39, no. 1, p. 42, 1953.

[3] M. Sion *et al.*, "On general minimax theorems.," *Pacific Journal of mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv:1312.6199 [cs]*, Feb. 2014. arXiv: 1312.6199.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv:1412.6572 [cs, stat]*, Mar. 2015. arXiv: 1412.6572.

[6] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," *arXiv:1511.04599 [cs]*, July 2016. arXiv: 1511.04599.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv:1706.06083 [cs, stat]*, Sept. 2019. arXiv: 1706.06083.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[9] A. Bemporad and M. Morari, "Robust model predictive control: A survey," in *Robustness in identification and control* (A. Garulli and A. Tesi, eds.), vol. 245, pp. 207–226, London: Springer London, 1999.

[10] L. Magni and R. Scattolini, "Robustness and robust design of mpc for nonlinear discrete-time systems," in *Assessment and future directions of nonlinear model predictive control*, pp. 239–254, Springer, 2007.

[11] D. A. Copp and J. P. Hespanha, “Simultaneous nonlinear model predictive control and state estimation,” *Automatica*, vol. 77, pp. 143–154, Mar. 2017.

[12] P. J. Huber, “Robust estimation of a location parameter,” in *Breakthroughs in statistics*, pp. 492–518, Springer, 1992.

[13] R. G. Staudte and S. J. Sheather, *Robust estimation and testing*, vol. 918. John Wiley & Sons, 2011.

[14] G. C. Calafiore and M. C. Campi, “The scenario approach to robust control design,” *IEEE Transactions on automatic control*, vol. 51, no. 5, pp. 742–753, 2006.

[15] C. Bandi and D. Bertsimas, “Tractable stochastic analysis in high dimensions via robust optimization,” *Mathematical Programming*, vol. 134, pp. 23–70, Aug. 2012.

[16] R. Chinchilla and J. P. Hespanha, “Optimization-based estimation of expected values with application to stochastic programming,” in *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 6356–6361, IEEE, 2019.

[17] A. Ben-Tal and A. Nemirovski, “Robust optimization - methodology and applications,” *Mathematical Programming*, vol. 92, pp. 453–480, May 2002.

[18] A. Mutapcic and S. Boyd, “Cutting-set methods for robust convex optimization with pessimizing oracles,” *Optimization Methods & Software*, vol. 24, no. 3, pp. 381–406, 2009.

[19] D. Bertsimas, D. B. Brown, and C. Caramanis, “Theory and Applications of Robust Optimization,” *SIAM Review*, vol. 53, pp. 464–501, Jan. 2011. Publisher: Society for Industrial and Applied Mathematics.

[20] D. Bertsimas, I. Dunning, and M. Lubin, “Reformulation versus cutting-planes for robust optimization,” *Computational Management Science*, vol. 13, no. 2, pp. 195–217, 2016.

[21] C. Jin, P. Netrapalli, and M. I. Jordan, “What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?,” feb 2019.

[22] J. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, “Learning with opponent-learning awareness,” in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130, 2018.

[23] A. Letcher, J. Foerster, D. Balduzzi, T. Rocktäschel, and S. Whiteson, “Stable opponent shaping in differentiable games,” *arXiv preprint arXiv:1811.08469*, 2018.

[24] A. Shapiro, “Differentiability Properties of Metric Projections onto Convex Sets,” *Journal of Optimization Theory and Applications*, vol. 169, pp. 953–964, June 2016.

[25] D. P. Bertsekas, *Nonlinear Programming.* Athena Scientific, 3rd ed., 2016.

Appendix - Proofs of Theorems

Proposition 1 (First order necessary condition) *Assume $f(\cdot)$ is continuously differentiable and (x^*, y^*) is a local minmax. Then, $\forall y \in \mathcal{Y}$, $(y - y^*)' \boldsymbol{\nabla}_y f(x^*, y^*) \leqslant 0$. Moreover, $\exists \delta_0 > 0$ such that $\forall x \in \mathcal{X} : \|x - x^*\| < \delta_0$, $(x - x^*)' \boldsymbol{\nabla}_x f(x^*, y^*) \geqslant 0$.* □

Proof. Starting with the max, fix any $y \in \mathcal{Y}$ and let us denote $p_y := (y - y^*)$. Because $\mathcal{Y}$ is convex, for any $\beta \in [0, 1]$, $y^* + \beta p_y \in \mathcal{Y}$. As (x^*, y^*) is a local maximum, there exist $\tilde{\beta} : \forall \beta \in [0, \tilde{\beta}]$ the following inequality holds

$$0 \geqslant \frac{f(x^*, y^* + \beta p_y) - f(x^*, y^*)}{\beta}$$

Because $f(\cdot)$ is continuously differentiable, according to the mean value Theorem, there exist $\bar{\beta} \in [0, \beta]$ such that the previous inequality is equivalent to

$$0 \geqslant \partial_y f(x^*, y^* + \bar{\beta} p_y)' p_y.$$

Taking the limit as β goes to 0 finishes the first part of the proof. Now for the min, take the δ_0 from the definition of local minmax, fix any $x \in \mathcal{X} : \|x - x^*\| < \delta_0$ and let us denote $p_x := (x - x^*)$. Because $\mathcal{X}$ is convex, for any $\alpha \in [0, 1]$, $x^* + \alpha p_x \in \mathcal{X}$ and $\|p_x\| < \delta_0$. Take the function $h(\cdot)$ from the definition of local minmax, and define the local optimum

$$p_y^*(\alpha p_x) = \underset{p_y : y^* + p_y \in \mathcal{Y}, \|p_y\| < h(\alpha p_x)}{\arg\max} \quad .$$

By the definition of $h(\cdot)$ we have that $p_y^*(\alpha p_x) \to 0$ as $\alpha \to 0$. Then, as (x^*, y^*) is a local minmax,

$$\begin{aligned}
0 \leqslant\ & \frac{f(x^* + \alpha p_x, y^* + p_y^*(\alpha p_x)) - f(x^*, y^*)}{\alpha} \\
=\ & \frac{f(x^* + \alpha p_x, y^* + p_y^*(\alpha p_x)) - f(x^*, y^*) + f(x^* +, y^* + p_y^*(\alpha p_x)) - f(x^* +, y^* + p_y^*(\alpha p_x))}{\alpha} \\
\leqslant\ & \frac{f(x^* + \alpha p_x, y^* + p_y^*(\alpha p_x)) - f(x^*, y^* + p_y^*(\alpha p_x))}{\alpha} \\
=\ & \partial_x f(x^* + \bar{\alpha} p_x, y^* + p_y^*(\alpha p_x))' p_x \quad \text{for some } \bar{\alpha} \in [0, \alpha]
\end{aligned}$$

taking the limit as α goes to zero finishes the proof. ■

Proposition 2 (Computing local directions) *If $f(x, y)$ is continuously differentiable with respect to y and $\hat{f}_x(x, y)$ is continuously differentiable with respect to x on a neighborhood around a point $(\tilde{x}, \tilde{y})$ which is not a stationary point, then there exist α_0 and β_0 such that $\forall \alpha \in (0, \alpha_0)$ and $\forall \beta \in (0, \beta_0)$ the functions $d_x(\tilde{x}, \tilde{y})$ and $d_y(\tilde{x}, \tilde{y})$ are full descent ascent directions.* □

Proof. Consider the equations

$$\begin{aligned}
f(x^{k+1}, y^{k+1}) &= \hat{f}(x^k, y^k) + \alpha d_x(x^k, y^k)' \boldsymbol{\nabla}_x \hat{f}(x^k, y^k) + o(\alpha) \\
f(x^{k+1}, y^{k+1}) &= f(x^{k+1}, y^k) + \beta d_y(x^{k+1}, y^k)' \boldsymbol{\nabla}_y f(x^{k+1}, y^k) + o(\beta)
\end{aligned}$$

where we use the fact that $f(x^{k+1}, y^{k+1}) = \hat{f}(x^{k+1}, y^k)$. As the functions are continuously differentiable, there exist α_0 and β_0 such that $\forall \alpha \in (0, \alpha_0)$ and $\forall \beta \in (0, \beta_0)$ the terms $o(\alpha)$ and $o(\beta)$ are dominated. From (6) and (8), we have that $d_y(x^k, y^k)' \boldsymbol{\nabla}_y f(x^k, y^k) \geqslant 0$ and $d_x(x^k, y^k)' \boldsymbol{\nabla}_x \hat{f}(x^k, y^k) \leqslant 0$. As at least either $d_y(x^k, y^k)' \boldsymbol{\nabla}_y f(x^k, y^k)$ or $d_x(x^k, y^k)' \boldsymbol{\nabla}_x \hat{f}(x^k, y^k)$ is non zero, otherwise (x^k, y^k) would be a stationary point, then they are full descent ascent directions. ■

Lemma 1 (Simultaneous descent ascent are gradient related) *The full descent ascent directions $d_x(x, y)$ and $d_y(x, y)$ are gradient related meaning that for any sequence $\{(x^k, y^k)\}$ that converges to a nonstationary point, then the corresponding sequence $\{(d_x(x^k, y^k), d_y(x^k, y^k))\}$ is bounded and satisfies*

$$\limsup_{k\to\infty} d_x(x^k, y^k)'\boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k) \leqslant 0$$
$$\liminf_{k\to\infty} d_y(x^{k+1}, y^k)'\boldsymbol{\nabla}_y f(x^{k+1}, y^k) \geqslant 0 \qquad \square$$

with at least one inequality holding strictly and where $\hat{f}_x(x, y)$ is defined in (7).

Proof. The proof is inspired in the proof of Prop 3.3.1 of [25].

Assume that $\{(x^k, y^k)\}$ converges to a non stationary point $(\tilde{x}, \tilde{y})$. We need to prove the following four equations

$$\limsup_{k\to\infty} \left\|d_x(x^k, y^k)\right\| < \infty \tag{10a}$$
$$\limsup_{k\to\infty} \left\|d_y(x^k, y^k)\right\| < \infty \tag{10b}$$
$$\liminf_{k\to\infty} d_y(x^{k+1}, y^k)'\boldsymbol{\nabla}_y f(x^{k+1}, y^k) \geqslant 0 \tag{10c}$$
$$\limsup_{k\to\infty} d_x(x^k, y^k)'\boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k) \leqslant 0 \tag{10d}$$

By continuity of the projection (see Prop. 1.1.4 in [25]) and the differential continuity of $\boldsymbol{\nabla}_x \hat{f}_x(x, y)$ and $\boldsymbol{\nabla}_y f(x, y)$

$$\lim_{k\to\infty} \left\|d_y(x^k, y^k)\right\| = \|d_y(\tilde{x}, \tilde{y})\| < \infty$$
$$\lim_{k\to\infty} \left\|d_x(x^k, y^k)\right\| = \|d_x(\tilde{x}, \tilde{y})\| < \infty$$

which proves (10b) and (10a). To prove (10c) and (10d), first remember the property that, for any continuously differentiable function $\phi(x)$ on a convex set $\mathcal{X}$, if x^* is a local minimum, then $\boldsymbol{\nabla}\phi(x^*)'(x - x^*) \geqslant 0\ \forall x \in \mathcal{X}$; there is an equivalent property for a local maximum. Applying these condition to (6) and (8) we obtain

$$\left(B(x^{k+1}, y^k)d_y(x^{k+1}, y^k) - \boldsymbol{\nabla}_y f(x^{k+1}, y^k)\right)'(\tilde{d}_y - d_y(x^{k+1}, y^k)) \geqslant 0 \quad \forall \tilde{d}_y : d_y + y^k \in \mathcal{Y}$$
$$\Big(A(x^k, y^k)d_x(x^k, y^k) + \boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k)\Big)'(\tilde{d}_x - d_x(x^k, y^k)) \geqslant 0 \quad \forall \tilde{d}_x : d_x + x^k \in \mathcal{X}$$

The above equations hold for $(\tilde{d}_x, \tilde{d}_y) = (0, 0)$ which yields

$$\boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k)'d_x(x^k, y^k) \leqslant -d_x(x^k, y^k)A(x^k, y^k)d_x(x^k, y^k) \leqslant -c_1\left\|d_x(x^k, y^k)\right\|^2 \tag{12a}$$
$$\boldsymbol{\nabla}_y f(x^{k+1}, y^k)'d_y(x^{k+1}, y^k) \geqslant d_y(x^{k+1}, y^k)'B(x^{k+1}, y^k)d_y(x^{k+1}, y^k) \geqslant c_3\left\|d_y(x^{k+1}, y^k)\right\|^2 \tag{12b}$$

where the last inequality is taken from the boundness of the eigenvalues of $A(x^k, y^k)$ and $B(x^{k+1}, y^k)$. Taking the limit we obtain

$$\liminf_{k\to\infty} d_y(x^{k+1}, y^k)'\boldsymbol{\nabla}_y f(x^{k+1}, y^k) \geqslant c_3\|d_y(\tilde{x}, \tilde{y})\|^2 \geqslant 0$$
$$\limsup_{k\to\infty} d_x(x^k, y^k)'\boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k) \leqslant -c_1\|d_x(\tilde{x}, \tilde{y})\|^2 \leqslant 0$$

with at least one inequality holding strictly because $(\tilde{x}, \tilde{y})$ is not a stationary point. ■

Theorem 1 (Constant step size) *Let $\{(x^k, y^k)\}$ be a full descent ascent sequence with $\alpha^k = \beta^k = 1$, $A(x, y) = A$ and $B(x, y) = B$. Assume that $\forall x_1, x_2 \in \mathcal{X}$ and $\forall y_1, y_2 \in \mathcal{Y}$ the exist constants $L_x, L_y > 0$ such that the following smoothness condition holds:*

$$\|\boldsymbol{\nabla}_x f(x_1, y_1) - \boldsymbol{\nabla}_x f(x_2, y_2)\| < L_x \sqrt{\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2}$$

$$\|\boldsymbol{\nabla}_y f(x_1, y_1) - \boldsymbol{\nabla}_y f(x_2, y_2)\| < L_y \sqrt{\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2}$$

If $2\Big(L_x\sqrt{1 + \lambda_{min}(B)\, L_y^2}\Big)^{-1} > \lambda_{min}(A)^{-1}$ and $2\, L_y^{-1} > \lambda_{min}(B)^{-1}$ then every limit point of $\{(x^k, y^k)\}$ is a stationary point of $f(\cdot)$. Moreover, $d_x(x, y)$ and $d_y(x, y)$ are full descent ascent directions, meaning that

$$f(x^{k+1}, y^{k+1}) \leqslant f(x^k, y^k + d_y(x^k, y^k))$$
$$f(x^{k+1}, y^{k+1}) \geqslant f(x^{k+1}, y^k)$$

with at least one of the inequalities holding strictly. □

Proof. Let us start proving the property for the max. Using the property known as the ascent lemma for Lipschitz function (see Prop. A.24 in [25]) we have that

$$f(x^{k+1}, y^k + d_y(x^k, y^k)) - f(x^{k+1}, y^k) \geqslant \boldsymbol{\nabla}_y f(x^{k+1}, y^k)' d_y(x^k, y^k) - \frac{L_y}{2}\big\|d_y(x^{k+1}, y^k)\big\|$$

Combining this result with (12b) where $c_3 := \lambda_{min}(B)$ we obtain

$$f(x^{k+1}, y^{k+1}) - f(x^{k+1}, y^k) \geqslant \left(\lambda_{min}(B) - \frac{L_y}{2}\right)\big\|d_y(x^{k+1}, y^k)\big\| \geqslant 0$$

where the right most inequalities hold because $\lambda_{min}(B) > L_y/2$. So if $(\bar{x}, \bar{y})$ is a limit point of a subsequences $\{(x^k, y^k)\}_{\mathcal{K}}$ then

$$\lim_{k\to\infty, k\in\mathcal{K}} f(x^{k+1}, y^{k+1}) - f(x^{k+1}, y^k) = 0$$

implying, by continuity of the projection, that $\|d_y(\bar{x}, \bar{y})\| = 0$.

For the min, take $\hat{f}_x(x, y)$ as defined in (7) and consider the following inequalities

$$\begin{aligned}
\Big\|\boldsymbol{\nabla}_x \hat{f}_x(x + d_x(x, y), y) - \boldsymbol{\nabla}_x \hat{f}_x(x, y)\Big\| &= \Big\|\boldsymbol{\nabla}_x f\Big(x + d_x(x, y), y + d_y\Big(x + d_x(x, y), y\Big)\Big) - \boldsymbol{\nabla}_x f(x, y + d_y(x, y))\Big\| \\
&\leqslant L_x\sqrt{\|d_x(x, y)\|^2 + \Big\|d_y\Big(x + d_x(x, y), y\Big) - d_y(x, y)\Big\|^2} \\
&\leqslant L_x\sqrt{\|d_x(x, y)\|^2 + \Big\|B^{-1}\boldsymbol{\nabla}_y f\Big(x + d_x(x, y), y\Big) - B^{-1}\boldsymbol{\nabla}_y f(x, y)\Big\|^2} \\
&\leqslant L_x\sqrt{\|d_x(x, y)\|^2 + \lambda_{min}(B)\Big\|\boldsymbol{\nabla}_y f\Big(x + d_x(x, y), y\Big) - \boldsymbol{\nabla}_y f(x, y)\Big\|^2} \\
&\leqslant L_x\sqrt{\|d_x(x, y)\|^2 + \lambda_{min}(B)\, L_y^2 \|d_x(x, y)\|^2} \\
&= L_x\sqrt{1 + \lambda_{min}(B)\, L_y^2}\|d_x(x, y)\|
\end{aligned}$$

where in the third line we used the fact that projections are nonexpansive (see Prop. 1.1.4 in [25]). These imply that the function $\hat{f}_x(x, y)$ is also smooth with constant $L_x\sqrt{1 + \lambda_{min}(B)\, L_y^2}$. So using the equivalent steps as for the max we arrive to

$$\hat{f}_x(x^k + d_x(x^k, y^k), y^k) - \hat{f}_x(x^k, y^k) \leqslant \left(\frac{L_x\sqrt{1 + \lambda_{min}(B)\, L_y^2}}{2} - \lambda_{min}(A)\right)\big\|d_x(x^k, y^k)\big\| \leqslant 0$$

where the right most equality hold because $\lambda_{min}(A) > L_x\sqrt{1+\lambda_{min}(B)\,L_y^2}\,/2$. So if $(\bar{x},\bar{y})$ is a limit point of a subsequences $\{(x^k,y^k)\}_{\mathcal{K}}$ then

$$\lim_{k\to\infty,k\in\mathcal{K}} \hat{f}_x(x^{k+1},y^k) - \hat{f}_x(x^k,y^k) = 0$$

implying, by continuity of the projection, that $\|d_x(\bar{x},\bar{y})\| = 0$. Therefore that $(\bar{x},\bar{y})$ is a stationary point. ■

Theorem 2 (Convergence of Armijo) *Every limit point of a sequence $\{(x^k,y^k)\}$ generated by Algorithm 1 is a stationary point.* □

Proof. This proof is inspired by the proof of Prop. 1.2.1 in [25]. Take $\hat{f}_x(x,y)$ as defined in (7) and, in order to have shorter expressions, let us define $d_x^k := d_x(x^k,y^k)$ and $d_y^k := d_y(x^{k+1},y^k)$.

As $f(\cdot)$ and $\hat{f}_x(\cdot)$ are continuous function, then as $(\bar{x},\bar{y})$ is a limit point of $\{(x^k,y^k)\}$ then $f(\bar{x},\bar{y})$ is a limit point of $\{f(x^k,y^k)\}$ and equivalent to $\hat{f}_x(\cdot)$. Moreover, $f(\bar{x},\bar{y})$ is also a limit point of $\{f(x^{k+1},y^k)\}$.

Starting with the max. From the previous argument, we have that

$$f(x^{k+1},y^k) - f(x^{k+1},y^{k+1}) \to 0.$$

By the choice of direction in (8) we have that $d_y^k{}'\boldsymbol{\nabla}_y f(x^{k+1},y^k) \geqslant 0$. Combining this with the Armijo rule in (9b) we have that

$$f(x^{k+1},y^k) - f(x^{k+1},y^{k+1}) \leqslant -\sigma_y\beta^k d_y^k{}'\boldsymbol{\nabla}_y f(x^{k+1},y^k) \leqslant 0 \tag{13}$$

Therefore we obtain that

$$\lim_{k\to\infty} \beta^k d_y^k{}'\boldsymbol{\nabla}_y f(x^{k+1},y^k) = 0 \tag{14}$$

Now the min. Combining (12b) and (14) implies that $\beta^k\, d_y^k \to \mathbf{0}_n$. And as $\hat{f}_x(\cdot)$ is continuous, we obtain

$$\hat{f}_x(x^k,y^k) - f(x^{k+1},y^{k+1}) \to 0.$$

By the choice of descent direction we have that $d_x^k{}'\boldsymbol{\nabla}_x\hat{f}_x(x^k,y^k) \leqslant 0$, and by the Armijo rule

$$\hat{f}_x(x^k,y^k) - f(x^{k+1},y^{k+1}) \geqslant -\sigma_x\,\alpha^k d_x^k{}'\boldsymbol{\nabla}_x\hat{f}_x(x^k,y^k) \geqslant 0$$

Therefore we obtain

$$\lim_{k\to\infty} \alpha^k d_x^k{}'\boldsymbol{\nabla}_x\hat{f}_x(x^k,y^k) = 0 \tag{15}$$

As d_x^k and d_y^k are gradient related from Lemma 1, in order for (14) and (15) to hold simultaneously either $(\bar{x},\bar{y})$ is a stationary point or $\alpha^k \to 0$ or $\beta^k \to 0$.

We will assume, in order to arrive to a contradiction, that $(\bar{x},\bar{y})$ is not a stationary. We will start by assuming that $\{\beta^k\} \to 0$, and show that it implies that $\{\alpha^k\} \to 0$ and then show it leads to a contradiction.

The core argument used to prove the contradiction relies in the following observation. If $\{\beta^k\} \to 0$ it means that there exist a $\bar{k}$, such that for each $k > \bar{k}$

$$f\big(x^k+\alpha^k d_x^k, y^k\big) - f\left(x^k+\alpha^k d_x^k, y^k + \frac{\beta}{r_y}^k d_y^k\right) > -\frac{\beta}{r_y}^k \sigma_y d_y^k{}'\boldsymbol{\nabla}_y f\big(x^k+\alpha^k d_x^k, y^k\big). \tag{16}$$

This equation holds because $\{\beta^k\} \to 0$ implies that the alternating backtracking algorithm will always need to run at least one time after some point, which we called $\bar{k}$. If the alternating backtracking algorithm ran at least one time it means that the Armijo conditions for the max was not verified for β^k/r_y, otherwise there would not have been the need to run another iteration of the backtracking, which justifies (16).

Since the search direction d_y^k is gradient related, then $\{d_y^k\}$ is bounded and so there exists a subsequences $\{d_y^k\}_{\bar{\mathcal{K}}}$ of $\{d_y^k\}$ such that $\{d_y^k\}_{\bar{\mathcal{K}}}$ converges to some point $\bar{d}_y$. Then, $\forall k \in \bar{\mathcal{K}}, k > \bar{k}$

$$\frac{f(x^k + \alpha^k d_x^k, y^k) - f(x^k + \alpha^k d_x^k, y^k + \beta^k/r_y d_y^k)}{\beta^k/r_y} > -\sigma_y d_y^k{}' \boldsymbol{\nabla}_y f(x^k + \alpha^k d_x^k, y^k) \tag{17}$$

By the mean value theorem, this relation can be written as

$$-d_y^k{}' \boldsymbol{\nabla}_y f(x^k + \alpha^k d_x^k, y^k + \tilde{\beta}^k d_y^k) > -\sigma_y d_y^k{}' \boldsymbol{\nabla}_y f(x^k + \alpha^k d_x^k, y^k) \tag{18}$$

with $\tilde{\beta}^k \in [0, \beta^k/r_y]$. Now taking the limit as $k \to \infty, k \in \bar{\mathcal{K}}$ and because $\{\beta^k\} \to 0$ we obtain

$$-\bar{d}_y{}' \boldsymbol{\nabla}_y f(\bar{x}, \bar{y}) \geqslant -\sigma_y \bar{d}_y{}' \boldsymbol{\nabla}_y f(\bar{x}, \bar{y}) \Leftrightarrow 0 \geqslant (1 - \sigma_y) \bar{d}_y{}' \boldsymbol{\nabla}_y f(\bar{x}, \bar{y}) \Rightarrow 0 \geqslant \bar{d}_y{}' \boldsymbol{\nabla}_y f(\bar{x}, \bar{y}).$$

There are two possible cases. The first one is that the last inequality holds strictly, *i.e.,* $\bar{d}_y{}' \boldsymbol{\nabla}_y f(\bar{x}, \bar{y}) < 0$. This contradicts the assumption that $\bar{d}_y$ is gradient related, therefore this case is not possible. The second case is that $\bar{d}_y{}' \boldsymbol{\nabla}_y f(\bar{x}, \bar{y}) = 0$. By contradiction assumption, $(\bar{x}, \bar{y})$ is not a stationary point, meaning $\lim_{k\to\infty} \alpha^k d_x^k{}' \boldsymbol{\nabla}_x \hat{f}_x((x^k, y^k) \neq 0$ (otherwise $(\bar{x}, \bar{y})$ is a stationary point). By (15) this implies that $\{\alpha^k\} \to 0$. Analogously to the previous case, if $\{\alpha^k\} \to 0$ then there exist a $\bar{k}$ such that for each $k > \bar{k}$

$$\hat{f}_x(x^k, y^k) - \hat{f}_x\left(x^k + \frac{\alpha^k}{r_x} d_x^k, y^k\right) < -\sigma_x \frac{\alpha^k}{r_x}^{k} d_x^k{}' \boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k). \tag{19}$$

Using equivalent arguments as above, we arrive to the conclusion that there is a subsequences $\{d_x^k\}_{\bar{\mathcal{K}}}$ that converges to some point $\bar{d}_x$ and that satisfies $0 \leqslant \bar{d}_x{}' \boldsymbol{\nabla}_x f(\bar{x}, \bar{y})$ which, if the inequality is strict, contradicts the assumption that d_x is gradient related , or contradicts the proof assumption that $(\bar{x}, \bar{y})$ is not a stationary point. Therefore, by contradiction $(\bar{x}, \bar{y})$ is a stationary point. ■

Theorem 3 (Capture Theorem) *Let $\{(x^k, y^k)\}$ be a sequence generated by the full descent ascent direction method using either the Theorem 1 or Theorem 2. Let (x^*, y^*) be an isolated local minmax on a neighborhood where it is also the only stationary point. Then there exist a neighborhood $S_x \subset \mathcal{X}$ around x^* and a neighborhood $S_y \subset \mathcal{Y}$ around y^* such that if for some $\bar{k}$, $(x^{\bar{k}}, y^{\bar{k}}) \in S_x \times S_y$ then* $\lim_{k, k > \bar{k}} (x^k, y^k) = (x^*, y^*)$. □

Proof. This proof is inspired by the proof of Prop. 1.2.4 of [25]. Let the interval $[0, \delta_0]$ and the function $h(\cdot)$ be the ones associated to the local minmax (x^*, y^*) according to Definition 1 . Take $\hat{f}_x(x, y)$ as defined in (7) and, in order to have shorter expressions, let us define $d_x^k := d_x(x^k, y^k)$ and $d_y^k := d_y(x^{k+1}, y^k)$.

Let us now take a specific $\delta \in [0, \delta_0]$. By definition, (x^*, y^*) is also a local minmax in that interval. Define for $t \in [0, \delta]$ and $t \in [0, h(\delta)]$ the functions

$$\begin{aligned} \phi_x(t, y) &= \min_{x \in \mathcal{X}: t \leqslant \|x^* - x\| \leqslant \delta} \hat{f}_x(x, y) - \hat{f}_x(x^*, y^*) \\ \phi_y(t, x) &= \max_{y \in \mathcal{Y}: t \leqslant \|y^* - y\| \leqslant h(\delta)} f(x, y) - f(x^*, y^*) \end{aligned}$$

For a fixed y, $\phi_x(t, y)$ is an increasing function of t and for a fixed x, $\phi_y(t, x)$ is a decreasing function of t. Given any $\epsilon_x \in (0, \delta]$ and $\epsilon_y \in (0, h(\delta)]$, take $r_x \in (0, \epsilon_x]$ and $r_y \in (0, \epsilon_y]$ such that

$$\|x - x^*\| < r_x \quad \Rightarrow \quad \|x - x^*\| + c_1^{-1} \left\| \boldsymbol{\nabla}_x \hat{f}_x(x, y) \right\| < \epsilon_x \tag{20a}$$

$$\|y - y^*\| < r_y \quad \Rightarrow \quad \|y - y^*\| + c_3^{-1} \|\boldsymbol{\nabla}_y f(x, y)\| < \epsilon_y \tag{20b}$$

where c_1 and c_3 are from Assumption 1. Consider the open sets

$$\begin{aligned} S_x &:= \{x \in \mathcal{X} : \|x - x^*\| < \epsilon_x \text{ and } \forall y : \|y - y^*\| < \epsilon_y,\ \hat{f}_x(x,y) - f(x^*,y^*) < \phi_x(r_x, y)\} \\ S_y &:= \{y \in \mathcal{Y} : \|y - y^*\| < \epsilon_y \text{ and } \forall x : \|x - x^*\| < \epsilon_x,\ f(x,y) - f(x^*,y^*) < \phi_y(r_y, x)\}. \end{aligned}$$

Now we prove that $x^k \in S_x \quad \Rightarrow \quad x^{k+1} \in S_x$ and that $y^k \in S_y \quad \Rightarrow \quad y^{k+1} \in S_y$. Starting with x^k, as $x^k \in S_x$ and $y^k \in S_y$, then

$$\phi_x\left(\left\|x^* - x^k\right\|, y^k\right) \leqslant \hat{f}_x(x^k, y^k) - f(x^*, y^*) < \phi_x(r_x, y^k)$$

where the right inequality derives from the definition of S_x and the left inequality from the definition of $\phi_x(\cdot)$. As $\phi_x(\cdot)$ is increasing in t, the previous relation implies $\left\|x^* - x^k\right\| < r_x$. Now we use the fact that in both Theorem 1 and 1 we have that $\alpha^k \leqslant 1$. Moreover, because the projection is a contracting map $\left\|d_x^k\right\| \leqslant c_1^{-1}\left\|\boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k)\right\|$ we obtain

$$\left\|x^{k+1} - x^*\right\| \leqslant \left\|x^k - x^*\right\| + \left\|\alpha^k d_x^k\right\| \leqslant \left\|x^k - x^*\right\| + c_1^{-1}\left\|\boldsymbol{\nabla}_x \hat{f}_x(x^k, y^k)\right\| \leqslant \epsilon_x$$

where the last inequality derives from (20a). Now looking back at the max from the previous equation we have that $\left\|x^{k+1} - x^*\right\| \leqslant \epsilon_x$ which implies

$$\phi_y\left(\left\|y^* - y^k\right\|, x^{k+1}\right) \geqslant f(x^{k+1}, y^k) - f(x^*, y^*) > \phi_y(r_y, x^{k+1}).$$

As $\phi_y(\cdot)$ is decreasing in t, the previous relation implies $\left\|y^* - y^k\right\| < r_y$. Now using the assumptions that $\beta^k \leqslant 1$ (same argument as α^k) and because the projection is a contracting map $\left\|d_y^k\right\| \leqslant c_3^{-1}\left\|\boldsymbol{\nabla}_x f(x^{k+1}, y^k)\right\|$ we obtain

$$\left\|y^{k+1} - y^*\right\| \leqslant \left\|y^k - y^*\right\| + \left\|\beta^k d_y^k\right\| \leqslant \left\|y^k - y^*\right\| + c_3^{-1}\left\|\boldsymbol{\nabla}_x f(x^{k+1}, y^k)\right\| \leqslant \epsilon_y$$

As $\{(x^k, y^k)\}$ is a full descent ascent sequence

$$\begin{cases} \hat{f}_x(x^{k+1}, y^k) - \hat{f}_x(x^*, y^*) \leqslant \hat{f}_x(x^k, y^k) - \hat{f}_x(x^*, y^*) < \phi_x(r_x, y^k) \\ \left\|x^{k+1} - x^*\right\| \leqslant \epsilon_x \\ \left\|y^{k+1} - y^*\right\| \leqslant \epsilon_y \end{cases} \quad \Rightarrow \quad x^{k+1} \in S_x$$

and

$$\begin{cases} f(x^{k+1}, y^{k+1}) - f(x^*, y^*) \leqslant f(x^{k+1}, y^k) - f(x^*, y^*) < \phi_y(r_y, x^{k+1}) \\ \left\|x^{k+1} - x^*\right\| \leqslant \epsilon_x \\ \left\|y^{k+1} - y^*\right\| \leqslant \epsilon_y \end{cases} \quad \Rightarrow \quad y^{k+1} \in S_y$$

Finally, by induction we have that if for some $\bar{k}$, $x^{\bar{k}} \in S_x$ and $y^{\bar{k}} \in S_y$, then $x^k \in S_x$ and $y^k \in S_y$ $\forall k > \bar{k}$. Let $\bar{S}_x$ and $\bar{S}_y$ be the closure of S_x and S_y. They are compact sets, therefore the sequence (x^k, y^k) must have at least one limit point which is a stationary point according to Theorem 1 and Theorem 2 . As the only stationary point is (x^*, y^*), therefore $(x^k, y^k) \to (x^*, y^*)$. ■