universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

’Quantitative convergence estimates of deterministic and stochastic methods for optimization and minimax problems’

verfasst von / submitted by

Axel Böhm

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2020 / Vienna 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 796 605 405
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Mathematik
Betreut von / Supervisor:	Univ.-Prof. Dr. Radu Ion Boţ
Mitbetreut von / Co-Supervisor:	ao. Univ.-Prof. Dr. Hermann Schichl

Abstract

Nonsmoothness plays a critical role in many optimization problems. Sometimes it is put into the model purposely to induce desirable properties in the solution, most notably sparsity, as it is the case with the composite models we study in the first half of this thesis. Although, the used nonsmooth functions tend to be simple, difficulty arises through the composition with another operator. We study such problems in a classical convex setting by proposing a randomized method and testing it on numerical experiments in image denoising and deblurring as well as completely positive matrix factorization. Additionally we propose a more sophisticated nonconvex formulation together with a novel method including convergence analysis for this setting. In either case, our approach is heavily inspired by a smoothing strategy via the *Moreau envelope*.

Other times the nonsmoothness originates naturally, for example due to the fact that the objective is derived from an auxiliary maximization problem. We study such *minimax* (a.k.a. saddle point) problems in the second half in a convex and nonconvex setting. While these types of problems also arise from two-player zero-sum games we emphasize applications in machine learning, in particular *generative adversarial networks (GANs)*. In the convex setting we propose a modification of Tseng's method while for the nonconvex problem we prove novel convergence rates for the well established *gradient descent ascent method (GDA)*.

In general we focus on *full splitting* methods which aim to tackle the nonsmoothness via the *proximal operator* and avoid convoluted inner loops or the need for subproblems. Similarly, only first-order information and preferably even only stochastic estimators of the involved gradients. These methods do not always achieve the best theoretical convergence rates but are nevertheless highly popular due to their simplicity and because they also tend to be very competitive in practice. For all presented methods we provide a rigorous analysis in terms of convergence rates.

Zusammenfassung

Nichtdifferenzierbarkeit spielt eine kritische Rolle in vielen verschiedenen Optimierungsproblemen. Manchmal wird sie künstlich hinzugefügt um wünschenswerte Eigenschaften in der Lösung zu erzeugen. Dies ist der Fall bei den Problemen denen wir uns in der ersten Hälfte dieser Arbeit widmen. Obwohl die involvierten nichtglatten Funktionen typischerweise simpel sind, entsteht die Schwierigkeit dadurch, dass sie mit einem anderen Operator hintereinander ausgeführt werden. Wir betrachten solche Probleme in einer klassischen konvexen Formulierung und stellen ein neues randomisiertes Verfahren vor, welches wir in numerischen Experimenten in der Bildverarbeitung und Matrixzerlegung testen. Zusätzlich stellen wir eine komplexere nichtkonvexe Version desselben Problems, gemeinsam mit einem neuen Verfahren, vor.

In anderen Fällen hingegen, entsteht Nichtdifferenzierbarkeit ganz natürlich, beispielsweise dadurch, dass die Zielfunktion einem inneren Maximierungsproblem entstammt. Wir behandeln solche Sattelpunktprobleme in der zweiten Hälfte der Arbeit. Solche Formulierungen haben ihren Ursprung, unter anderem in Nullsummenspielen zweier konkurrierender Parteien. Wir hingegen legen besonderes Augenmerk auf Anwendungen im Maschinellen Lernen, insbesondere sogenannte *generative adversarial networks (GANs).* Im konvexen Fall stellen wir eine Modifikation von dem bekannten Verfahren von Tseng vor, während wir im Nichtkonvexen ein simples Gradientenverfahren analysieren.

Im Allgemeinen konzentrieren wir uns auf Splitting-Methoden, die sich dadurch auszeichnen, dass die Nichtglattheit mittels des Proximalpunktoperators behandelt wird und aufwendige Subroutinen vermieden werden. Diese Verfahren erreichen zwar nicht immer die besten theoretischen Konvergenzraten, sind aber dennoch sehr beliebt aufgrund ihrer Einfachheit und Kompetitivität in praxisrelevanten Anwendungen.

Contents

1 Introduction

Nonsmoothness plays a critical role in many optimization problems. Sometimes it is deliberately put into a model to induce desirable properties in the solution, most notably sparsity. Typically this is done by adding a 1-norm of the decision variable to the objective. In linear regression the resulting problem is known as Lasso [100] and through the imposed sparsity irrelevant features can be excluded more easily. Similar applications can be found, for example, in signal processing [81]. In inverse problems, particularly image processing, usually sparsity is not sought in individual pixels but in the difference of neighboring pixels. This consideration results in a problem formulation where the 1-norm is composed with a discretized gradient a.k.a. total variation regularization [94].

Other times the nonsmoothness is more intrinsic and originates, for example, from the fact that the objective function itself is given as the solution of a maximization problem, see Chapter 5 and 6. Such minimax problems arise in various applications such as zero-sum games in the sense of game theory [104]. More recently they attracted increased interest due to their application in different machine learning tasks such as robust adversarial learning [98], learning with uncertain data [31], multi-agent reinforcement learning [79], learning with decomposable losses [39, 107] and the training of generative adversarial networks (GANs).

(Near) optimality. In the remainder we will pool convex-concave minimax problems and convex single-objective problems together for they similar proeperties and refer to them as just *convex*. Defining optimality for such problems is elementary. Typically one is looking for a global minimum in the case of single-objective optimization or a saddle point, see (6.2), for minimax problems. In either way these notions can be equivalently characterized by a first order condition. For nonconvex problems one can usually not expect to find such global solutions and is typically content with finding stationary points. Even for nonsmooth nonconvex problems an appropriate generalization of vanishing subgradients, see Definition 2.3.2, is straight forward. However, since we are generally interested in convergence rates we have to quantify how close a given point is to stationarity. For smooth functions this can be measured via the norm of the gradient. For nonsmooth functions such an approach fails even in the convex case. Consider, for example, the absolute value function. Every point different from the solution, no matter how close, will have (sub)gradients bounded away from zero. This somewhat troublesome observation can be remedied by measuring criticality in terms of closeness to a point with a small subgradient. In the weakly convex setting the gradient of the *Moreau envelope*, see Definition 2.3.6, captures this property and additionally provides a framework [35,36] for studying convergence.

Solution methods. For solving nonsmooth problems a natural first approach would be to seek to devise an appropriate generalized notion of a gradient, see Definition 2.1.7 and 2.3.2, and then continue to use smooth first-order methods. It is easy to see that subgradient descent with fixed stepsize fails to converge even on the simplest cases, such as the absolute value function. However, by employing more sophisticated stepsize regimes, this undesirable behavior can be circumvented [35, 72], but this typically results in slow methods. This approach also disregards the fact that nondifferentiability is often given by a simple algebraic description of the functions involved. By making use of the *proximal operator*, which inherently relies on the fact that the nonsmoothness arises in a structured way, we can devise faster, more problem adapted methods [9, 11, 28, 37]. As seen by the above mentioned applications in imaging, inverse problems or machine learning, we generally deal with problems which possibly exhibit a large number of variables but usually do not require high precision in the solutions. For this reason we aim to devise methods which rely on first-order information and the proximal operator only. To further cope with the possibly large scale of the problems we further emphasis the use of stochastic methods which only require samples of the objective function [14,36,89].

1.1 Overview

Chapter 2 is devoted to establishing notation and introducing the relevant preliminary concepts and statements. We will recall basic elements from convex and nonsmooth analysis in order to provide a compelling and self-contained reading experience. Most notably, the Moreau envelope which will play an intricate role in the forthcoming Chapter 3, 4 and 6, is introduced.

Composition with a linear operator. The first half of the main body is devoted to problems of the type

$$\min_x f(x) + g(Ax) \tag{L}$$

for a nonsmooth function g, composed with the linear operator A. For its desirable properties, we want to make use of the proximal operator of g, see Definition 2.3.6, but avoid the one of $g \circ A$, as there is typically no formula available for the latter.

Chapter 3 is concerned with the convex version of (L). Instead of the, by now classical, approach via primal-dual methods [27, 28, 32, 105] we make use of an acceleration techniques [9, 73] similarly to [101] with respect to the smoothed version. The main convergence results are summarized in Theorem 3.3.2 and 3.4.4 for a deterministic and stochastic problem formulation, proving complexity bounds of $\mathcal{O}(\epsilon^{-1})$ and $\tilde{\mathcal{O}}(\epsilon^{-2})$, respectively ($\tilde{\mathcal{O}}$ hides logarithmic terms). We finish this chapter with numerical experiments in image denoising/deblurring, where $g \circ A$ corresponds to the total variation regularization [94] (the ROF model) as well as sparse completely positive matrix factorization where we use our method as subroutine for the prox-linear method [37]. This chapter is based on the article [15].

Chapter 4 on the other hand deals with the setting where problem (L) is assumed to be weakly convex. This allows us to study for example sparsity inducing regularizers which do not induce a bias, see Section 4.1.1. We propose a simple novel method based on vanilla gradient descent for the smoothed problem. While in the convex setting the methods proposed in Chapter 3 competes with the well known primal-dual methods, in the nonconvex setting there is no equivalent notion of duality. The convergence statement for the basic version of our proposed method is analyzed in Theorem 4.2.2, whereas the result for the more sophisticated extension which ensured improved feasibility guarantees is stated in Theorem 4.2.5. Overall we prove a convergence rate of $\mathcal{O}(\epsilon^{-3})$ which interpolates nicely between the optimal rate for smooth nonconvex problem of $\mathcal{O}(\epsilon^{-2})$ and the black box subgradient method [35] with no additional knowledge about the nonsmoothness requiring $\mathcal{O}(\epsilon^{-4})$ iterations. The result is also in line with other methods dealing with nonconvex problems where it is assumed that the nonsmoothness arises from some particular algebraic description of the problem such as the composition of a nonsmooth convex function with a smooth vector-valued function [37]; or due to the inner maximization of a saddle point problem [61, 99], see also Chapter 6. The article [13] is the basis of Chapter 4.

Minimax problems. The second half of this thesis studies so-called minimax (or saddle point) problems of the type

$$\min_x \max_y f(x) + \Phi(x, y) - h(x) \tag{SP}$$

for a differentiable function Φ and nonsmooth, convex regularizers f, h. The implicit assumption here is that neither the maximization nor the minimization can be solved in closed form, and only steps based on first-order information can be taken in either direction.

In **Chapter 5** we tackle the case where the coupling function Φ is convex-concave by using the well known forward-backward-forward method by Tseng [103] and prove novel convergence rates in the case of stochastic gradient evaluations. We also propose a modification which recycles old gradients but turns out to be a known scheme [66] related to *optimistic* gradient descent ascent [33, 34, 58]. We analyze both methods in a unified way for different stepsize choices in Theorem 5.3.9 and 5.3.13 proving a convergence rate of $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(\epsilon^{-2})$ for the deterministic and stochastic setting, respectively. We conclude this chapter, which is based on the article [12], with numerical experiments in GAN training.

In **Chapter 6** we turn to weakly convex-(strongly) concave saddle point problems. Under these assumptions the inner maximization problem yields a weakly convex function in the remaining variable x, which keeps the problem tractable. Most of the existing literature has focused on inner loop methods [61, 78, 86, 99, 110] which will either repeatedly approximate the inner maximization or regularize the problem by adding a quadratic proximity term and then repeatedly solve the resulting convex-concave saddle point problem. While these methods obtain the best convergence rates for this class of problems, in practice, single loop methods, such as *gradient descent ascent (GDA)*, are still highly popular [4, 41, 49]. In particular, we prove the first convergence rates for stochastic *alternating* GDA outside a convex-concave setting. We show $\mathcal{O}(\epsilon^{-4})$ and $\mathcal{O}(\epsilon^{-8})$ for weakly convex-strongly concave (Theorem 6.4.9) and weakly convex-concave problems (Theorem 6.3.13), respectively. We furthermore close the a gap in the deterministic case where [106] studied criticality of Φ and [60] analyzes simultaneous GDA. This chapter is based on [16].

The connection between (L) **and** (SP)**.** We also want to point out that problems of type (L) can be seen as a purely primal version of a saddle point problem where the coupling function is bilinear. In fact, if g is convex, then (L) can be solved by considering instead the minimax problem

$$\min_x \max_y f(x) + \langle Ax, y \rangle - g^*(y),$$

where g^* denotes the Fenchel conjugate of g. Similarly, the minimax problem (SP) can be seen as *just* a minimization task, where the objective function exhibits a max-structure.

1.2 Acknowledgements

I want to take the time and thank the people who contributed in some way or another towards the development of this dissertation. First and foremost, my gratitude goes to my advisor Professor Radu Ioan Boţ for his truly exceptional effort. Effort, not only in supervising me or his other students but also his dedication to accelerate the entire field of optimization and even all of mathematics. Never have I seen anybody equate their own success so much with the success of the people around them. Needless to say, I feel very fortunate to have him as my supervisor.

Likewise, I would like to thank my co-supervisor Hermann Schichl who played a crucial role in me pursuing this path. My gratefulness goes also out to the other members of the graduate school on computational optimization — professors and students alike for creating a sense of community, not easily matched amongst graduate programs. I'll also gladly mention the colleagues of our research group: Robert, Sorin, Sebastian, Dennis, Khoa who would always lend an ear to whatever was on my mind at the time and have helpful comments for me; especially Michael through whom I experienced a new level of collaboration. Similarly, I was blessed to be part of the Vienna School of Mathematics whos members made this journey even more worthwile.

I also want to express my deep gratitude toward Aris Daniilidis and Steve Wright, who both welcomed me in their respective research groups for three month research visits each; enriching my life not only from an academic perspective.

Next, I would like to thank the two reviewers of this thesis, Simon Lacoste-Julien and Peter Richtárik, who took a significant amount of time of their busy days and agreed immediately to read this work.

Last but not least, I want to thank my family, friends and Hannah who not only supported me in so many different ways but also made life outside university so much more enjoyable.

My research was financially supported by the doctoral programme *Vienna Graduate School on Computational Optimization (VGSCO)*, FWF (Austrian Science Fund), project W 1260.

2 Preliminaries

We denote by $\mathbb{R}^d$ the d-dimensional Euclidean space, with its inner product by $\langle\cdot,\cdot\rangle$ and the generated norm by $\|\cdot\|$, where $\|x\|^2 = \langle x, x\rangle$ for $x \in \mathbb{R}^d$.

2.1 Convex analysis

Definition 2.1.1. The normal cone of the nonempty and convex set $C \subset \mathbb{R}^d$ is given by

$$N_C(x) = \{v \in \mathbb{R}^d : \langle v, u - x\rangle \leq 0 \quad \forall u \in C\},$$

for $x \in C$ and $N_C(x) = \emptyset$ for $x \notin C$.

Lemma 2.1.2. *For $\alpha \in \mathbb{R}$ and every $x, u \in \mathbb{R}^d$ we have that*

$$(1-\alpha)\|x-u\|^2 + \alpha\|u\|^2 \geq \alpha(1-\alpha)\|x\|^2.$$

Proof. See [6, Corollary 2.14]. □

Definition 2.1.3. We say that an operator between to Hilbert spaces $\mathbb{R}^d, \mathbb{R}^n$ is said to be *Lipschitz* with constant L, or L-Lipschitz, if

$$\|A(x) - A(u)\|_{\mathbb{R}^n} \leq L\|x-u\|_{\mathbb{R}^d}.$$

2.1.1 Convex functions

Definition 2.1.4. For an extended real valued function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ we introduce the following notions.

(i) The *domain* of g, denoted by $\operatorname{dom} g$, is defined as

$$\operatorname{dom} g := \{x \in \mathbb{R}^d : g(x) \neq +\infty\}.$$

(ii) We say that g is *proper* if its domain is nonempty.

(iii) We say that g is *convex* if for all $x, u \in \mathbb{R}^d$ and $0 \leq \alpha \leq 1$

$$g(\alpha x + (1-\alpha)u) \leq \alpha g(x) + (1-\alpha) g(u).$$

(iv) We say that g is *lower semicontinuous* if for all $x \in \mathbb{R}^d$

$$\liminf_{u \to x} g(u) \geq g(x).$$

Example 2.1.5. *The indicator function δ_C of a set $C \subseteq \mathbb{R}^d$ is defined as*

$$\delta_C(z) := \begin{cases} 0, & z \in C \\ +\infty, & \textit{otherwise.} \end{cases}$$

If the set C is nonempty, convex and closed, then δ_C is proper, convex and lower semicontinuous.

Definition 2.1.6 (Fenchel conjugate)**.** For a proper, convex and lower semicontinuous function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its *Fenchel conjugate* is denoted by g^* defined as a function from $\mathbb{R}^d$ to $\mathbb{R} \cup \{+\infty\}$, given by

$$g^*(p) := \sup_{x \in \mathbb{R}^d} \Big\{ \langle p, x \rangle - g(x) \Big\} \quad \forall p \in \mathbb{R}^d.$$

Definition 2.1.7 (Subdifferential)**.** For a function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ the *convex subdifferential* is given by

$$\partial g(x) := \Big\{ v \in \mathbb{R}^d : \langle v, u - x \rangle + g(x) \leq g(u) \quad \forall u \in \mathbb{R}^d \Big\}$$

for points $x \in \mathbb{R}^d$ where $g(x)$ is finite and the empty set otherwise. We call any element of this set a *subgradient* of g at x.

The (convex) subdifferential consists of all affine underestimates that touch the function at the given point.

Proposition 2.1.8 (Moreau decomposition)**.** *For a proper, convex and lower semicontinuous function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and its Fenchel conjugate g^* it holds that for all $\gamma > 0$*

$$x = \operatorname{prox}_{\gamma g}(x) + \gamma \operatorname{prox}_{g^*/\gamma}(x/\gamma) \quad \forall x \in \mathbb{R}^d.$$

Proof. See [6, Theorem 14.3 (ii)] □

In particular, if we have an analytic formula for the proximal operator of g, we also have a formula for the proximal operator of g^* and vice-versa.

Lemma 2.1.9 (Fermat's rule)**.** *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function. Then, $x^* \in \mathbb{R}^d$ is a minimizer of g, i.e. $g(x^*) \leq g(x)$ for all $x \in \mathbb{R}^d$, if and only if $0 \in \partial g(x^*)$.*

Proof. This follows immediately from the definition of the convex subdifferential. □

Definition 2.1.10 (Strong convexity)**.** For some $\mu > 0$, we say that

$$g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\} \text{ is } \mu\textit{-strongly convex} \text{ if } g - (\mu/2)\|\cdot\|^2 \text{ is convex.}$$

Lemma 2.1.11. *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper, μ-strongly convex and lower semicontinuous. Then g has a unique minimizer x^* and*

$$g(x^*) + \frac{\mu}{2}\|x - x^*\|^2 \leq g(x) \quad \forall x \in \mathbb{R}^d.$$

2.2 Differentiable functions

The class of differentiable functions with L-Lipschitz continuous gradient plays an important role in optimization. For short we call them L-smooth.

Lemma 2.2.1 (Descent lemma)**.** *For an L-smooth function $h : \mathbb{R}^d \to \mathbb{R}$ it holds that*

$$h(u) \leq h(x) + \langle \nabla h(x), u - x \rangle + \frac{L}{2}\|u - x\|^2 \quad \forall x, u \in \mathbb{R}^d.$$

Proof. See e.g. [6, Theorem 18.15]. □

The descent lemma states that every L-smooth function can be upper bounded by a quadratic function. By going from g to $-g$ we can see that the same statement holds for a lower bound.

The following lemma is a standard result for convex differentiable functions.

Lemma 2.2.2. *For a convex and L-smooth function $h : \mathbb{R}^d \to \mathbb{R}$ we have that*

$$h(x) + \langle \nabla h(x), u - x \rangle \leq h(u) - \frac{1}{2L}\|\nabla h(x) - \nabla h(u)\|^2 \quad \forall x, u \in \mathbb{R}^d.$$

Proof. See [77, Theorem 2.1.5]. □

The previous lemma strengthens the obvious inequality we would have deduced from the fact that the gradient is a subgradient of h which would yield that

$$h(x) + \langle \nabla h(x), u - x \rangle \leq h(u) \quad \forall x, u \in \mathbb{R}^d.$$

Lemma 2.2.3. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth and μ-strongly convex function. Then*

$$\langle \nabla h(x) - \nabla h(u), x - u \rangle \geq \frac{\mu L}{\mu + L}\|x - u\|^2 + \frac{1}{\mu + L}\|\nabla h(x) - \nabla h(u)\|^2.$$

If $\mu = 0$ the inequality still holds true and is known as cocoercivity*.*

Proof. See [77, Theorem 2.1.11] and [6, Theorem 18.15 (v)]. □

2.3 Weak convexity

Definition 2.3.1. For some $\rho \geq 0$, we say that

$g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *ρ-weakly convex if $g + (\rho/2)\|\cdot\|^2$ is convex.*

Weakly convex functions share some desirable properties with convex functions, but include many interesting nonconvex cases; see Section 4.1.1.

The concept of subgradient of a convex function can be adapted to weakly convex functions via the following definition.

Definition 2.3.2 (Fréchet subdifferential)**.** Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a function and x a point such that $g(x)$ is finite. Then, the *Fréchet subdifferential* of g at x, denoted by $\partial g(x)$, is the set of all vectors $v \in \mathbb{R}^d$ such that

$$g(u) \geq g(x) + \langle v, u - x \rangle + o(\|u - x\|) \quad \text{as } u \to x.$$

Lemma 2.3.3. *A lower semicontinuous function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is ρ-weakly convex if and only if for all Fréchet subgradients $v \in \partial g(x)$ the subgradient inequality holds:*

$$g(x) + \langle v, u - x \rangle - \frac{\rho}{2}\|u - x\|^2 \leq g(u) \quad \forall x, u \in \mathbb{R}^d.$$

Proof. See [36, Lemma 2.1] for a proof and more equivalent notions. The idea of the proof is to the definition of the convex subdifferential of $\varphi = g + (\rho/2)\|\cdot\|^2$ and the calculus rule that $\partial\varphi(x) = \partial g(x) + \rho x$, see [91, Exercise 8.8]. □

The previous lemma together with the descent lemma immediately implies that every L-smooth function is L-weakly convex.

Lemma 2.3.4. *If g is differentiable at the point $x \in \mathbb{R}^d$ then its Fréchet subdifferential consists of just the gradient $\partial g(x) = \{\nabla g(x)\}$.*

Proof. See [91, Excercise 8.8]. □

While the next result is standard for the gradient and convex subgradients we explicitly mention the general case.

Lemma 2.3.5. *For an L_g-Lipschitz continuous function $g : \mathbb{R}^d \to \mathbb{R}$ every Fréchet subgradient is bounded in norm by L_g.*

Proof. See [69, Theorem 3.52]. □

2.3.1 The Moreau envelope

Definition 2.3.6. For a proper, ρ-weakly convex and lower semicontinuous function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the *Moreau envelope* of g with the parameter $\lambda \in (0, \rho^{-1})$ is the function from $\mathbb{R}^d$ to $\mathbb{R}$ defined by

$$g_\lambda(x) := \inf_{u \in \mathbb{R}^d} \Big\{ g(u) + \frac{1}{2\lambda}\|u - x\|^2 \Big\}.$$

The *proximal operator* of the function λg is the arg min of the right-hand side in this definition, that is,

$$\operatorname{prox}_{\lambda g}(x) := \operatorname*{arg\,min}_{u \in \mathbb{R}^d} \Big\{ g(u) + \frac{1}{2\lambda}\|u - x\|^2 \Big\}. \tag{2.1}$$

Note that $\operatorname{prox}_{\lambda g}(x)$ is uniquely defined by (2.1) because the function being minimized is proper, lower semicontinuous and strongly convex. If g is in fact convex, i.e. $\rho = 0$, then λ can be chosen in $(0, +\infty)$.

Lemma 2.3.7. *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, ρ-weakly convex and lower semicontinuous function, and let $\lambda \in (0, \rho^{-1})$. Then the Moreau envelope $g_\lambda(\cdot)$ is continuously differentiable on $\mathbb{R}^d$ with gradient*

$$\nabla g_\lambda(x) = \frac{1}{\lambda}\left(x - \operatorname{prox}_{\lambda g}(x)\right) \quad \textit{for all } x \in \mathbb{R}^d.$$

This gradient is $\max\left\{\lambda^{-1}, \frac{\rho}{1-\rho\lambda}\right\}$-Lipschitz continuous. In particular, a gradient step with respect to the Moreau envelope corresponds to a proximal step, that is,

$$x - \lambda \nabla g_\lambda(x) = \operatorname{prox}_{\lambda g}(x), \quad \textit{for all } x \in \mathbb{R}^d. \tag{2.2}$$

Additionally, if g is convex, then g_λ is convex as well and the gradient of the Moreau envelope can also be characterized in terms of the proximal operator of the conjugate

$$\nabla g_\lambda = \operatorname{prox}_{\frac{1}{\lambda}g^*}(\cdot/\lambda).$$

Proof. For a proof of the first statement see [50, Corollary 3.4]. The statement for convex g can be found in [6, Proposition 12.29], but follows immediately from the Moreau decomposition, see Proposition 2.1.8. □

Lemma 2.3.7 not only clarifies the smoothness of the Moreau envelope, but also gives a way of computing its gradient via the proximal operator. Obviously, a smooth representation whose gradient could not be computed would be of only limited usefulness from an algorithmic standpoint. The only difference between the weakly convex and convex settings is that the Moreau envelope need not be convex in the former case.

Lemma 2.3.8. *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, ρ-weakly convex, and lower semicontinuous function, and let $\lambda \in (0, \rho^{-1})$. Then,*

$$\nabla g_\lambda(x) \in \partial g(\operatorname{prox}_{\lambda g}(x)) \quad \forall x \in \mathbb{R}^d. \tag{2.3}$$

Proof. From Definition 2.3.6, we have that

$$0 \in \partial g(\operatorname{prox}_{\lambda g}(x)) + \frac{1}{\lambda}(\operatorname{prox}_{\lambda g}(x) - x),$$

from which the result follows when we use (2.2). □

Lemma 2.3.9. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a ρ-weakly convex function that is L_g-Lipschitz continuous, and let $\lambda \in (0, \rho^{-1})$. Then the Moreau envelope g_λ is L_g-Lipschitz continuous as well*

$$|g_\lambda(x) - g_\lambda(u)| \leq L_g \|x - u\| \quad \forall x, u \in \mathbb{R}^d. \tag{2.4}$$

Therefore, $\|\nabla g_\lambda(x)\| \leq L_g$ and in particular

$$\|x - \operatorname{prox}_{\lambda g}(x)\| \leq \lambda L_g \quad \forall x \in \mathbb{R}^d. \tag{2.5}$$

Proof. By (2.3), we have for all $x \in \mathbb{R}^d$

$$\|\nabla g_\lambda(x)\| \leq \sup \left\{ \|v\| : \, v \in \partial g(\operatorname{prox}_{\lambda g}(x)) \right\} \leq L_g,$$

where we used Lemma 2.3.5 in the second inequality, which lets us conclude (2.4). The bound (2.5) follows immediately by considering the fact that $x - \operatorname{prox}_{\lambda g}(x) = \lambda \nabla g_\lambda(x)$ from Lemma 2.3.7. □

(The above two lemmata are proved for the case of g convex in [37, Lemma 2.1], with essentially the same proof.)

2.4 Stochastics

We want to recall basic notions from measure and probability theory which can be found in any introductory book on this matter.

Definition 2.4.1. For a set Ω and a sigma algebra $\mathcal{A}$ on this set we call the tuple $(\Omega, \mathcal{A})$ a *measurable space.*

Definition 2.4.2. If a measurable space $(\Omega, \mathcal{A})$ is additionally equipped with a probability measure $\mathbb{P}$ we call the triple $(\Omega, \mathcal{A}, \mathbb{P})$, a probability space.

A measurable mapping from a probability space to a measurable space is called a random variable. Usually when talking about random variables we will omit the spaces and sometimes even the probability measure, e.g. when talking about the expectation $\mathbb{E}[X]$ of the random variable X.

Definition 2.4.3 (Expectation)**.** For a random variable $X : \Omega :\to \mathbb{R}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ the *expected value* is defined as $\mathbb{E}[X] := \int_\Omega X(\omega) \, \mathrm{d}\mathbb{P}(\omega)$.

Definition 2.4.4 (Conditional expectation)**.** For a random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with finite expectation, its *conditional expectation* with respect to a subsigmaalgebra $\mathcal{S}$ of $\mathcal{A}$ denoted by $\mathbb{E}[X \,|\, \mathcal{S}]$, is the $\mathcal{S}$ measurable random variable fulfilling

$$\int_A X \, \mathrm{d}\mathbb{P} = \int_A \mathbb{E}[X \,|\, \mathcal{S}] \, \mathrm{d}\mathbb{P} \quad \forall A \in \mathcal{S}.$$

Although this definition is not constructive it is a standard task in measure theory to show that such a random variable exists and is unique (in an almost sure sense). We will regularly use the notation $\mathbb{E}[X \,|\, Y]$ for two random variables where we mean the conditional expectation of X with respect to the sigmaalgebra generated by Y.

Lemma 2.4.5. *Let the assumptions of Definition 2.4.4 hold.*

(i) If X is measurable with respect to $\mathcal{S}$, then $\mathbb{E}[X \,|\, \mathcal{S}] = X$.

(ii) If X is independent of $\mathcal{S}$, then $\mathbb{E}[X \,|\, \mathcal{S}] = \mathbb{E}[X]$.

The combination of the above properties culminates in the following statement.

Lemma 2.4.6 (Independence lemma)**.** *Let g be a function of two arguments such that*

$$\varphi(x) := \mathbb{E}[g(x, Y)].$$

If X and Y are two independent random variables, then

$$\mathbb{E}[g(X, Y) \mid X] = \varphi(X).$$

Proof. See [97, Lemma 2.5.3]. □

In the stochastic settings of Chapter 3, 5 and 6. We will often deal with the case where the objective function $F : \mathbb{R}^d \to \mathbb{R}$ is given for all $x \in \mathbb{R}^d$ as $\mathbb{E}_\xi[F(x, ; \xi)]$ for a random variable ξ (with a slight abuse of notation). We write $\mathbb{E}_\xi$ to emphasize that ξ is stochastic and not x, but leave it out later on. We will typically assume that the gradient of $F(x; \xi)$ with respect to x is an unbiased estimator for the gradient of F, i.e. that for all $x \in \mathbb{R}^d$

$$\mathbb{E}_\xi[\nabla F(x; \xi)] = \nabla F(x).$$

In order to analyze algorithms, which make use of such stochastic gradients, the iterates $(x_k)_{k \geq 0}$ will turn to be stochastic themselves. Thus, by using Lemma 2.4.6 we get that

$$\mathbb{E}[\nabla F(x_k; \xi) \mid x_k] = \nabla F(x_k)$$

as long as ξ is independent of x_k.

Lemma 2.4.7 (Tower property)**.** *Let $\mathcal{A}$ and $\mathcal{A}'$ be two sigma algebras such that $\mathcal{A} \subseteq \mathcal{A}'$. Then,*

$$\mathbb{E}\big[\mathbb{E}\big[X \mid \mathcal{A}'\big] \mid \mathcal{A}\big] = \mathbb{E}[X \mid \mathcal{A}].$$

In particular, the above lemmata implies the *law of total expectation* stating that

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{A}]] = \mathbb{E}[X]$$

for any sigma algebra $\mathcal{A}$.

3 Variable smoothing for convex composite problems

We aim to solve a structured convex optimization problem, where a nonsmooth function is composed with a linear operator. When opting for full splitting schemes, usually, primal-dual type methods are employed as they are effective and also well studied. However, under the additional assumption of Lipschitz continuity of the nonsmooth function which is composed with the linear operator we can derive novel algorithms through regularization via the Moreau envelope. Furthermore, we tackle large scale problems by means of stochastic oracle calls, very similar to stochastic gradient techniques. Applications to total variational denoising and deblurring are provided.

3.1 Problem setting

The problem at hand is the following structured convex optimization problem

$$\min_{x \in \mathbb{R}^d} \Big\{ F(x) := f(x) + g(Ax) \Big\}, \tag{3.1}$$

for $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ a proper, convex and lower semicontinuous function, $g : \mathbb{R}^n \to \mathbb{R}$ a, possibly nonsmooth, convex and L_g-Lipschitz continuous ($L_g > 0$) function, and $A : \mathbb{R}^d \to \mathbb{R}^n$ a nonzero linear operator.

Our aim will be to devise an algorithm for solving (3.1) following the *full splitting* paradigm (see [17, 18, 20, 21, 28, 32, 105]). In other words, we allow only proximal evaluations for simple nonsmooth functions, but no proximal evaluations for compositions with linear continuous operators, like, for instance, for $g \circ A$.

We will accomplish this feat by the means of the Moreau envelope, see Definition 2.3.6. The approach can be described as follows: we "smooth" g, i.e. we replace it by its Moreau envelope, and solve the resulting optimization problem by an *accelerated proximal gradient algorithm* (see [9, 26, 73]). This approach is similar to those in [19, 22, 23, 74, 76], where a convergence rate of $\tilde{\mathcal{O}}(1/k)$ is proved. However, our techniques (for the deterministic case) resemble more the ones in [101], where an improved rate of $\mathcal{O}(1/k)$ is shown. The most notable difference between this work and ours being the fact that we use a simpler stepsize and treat the stochastic case.

The only other family of methods able to solve problems of type (3.1) are the so called primal-dual algorithms, first and foremost the *primal-dual hybrid gradient (PDHG)* introduced in [28]. In comparison, this method does not need the Lipschitz continuity of g in order to prove convergence. However, in this very general case, convergence rates can only be shown for the so-called *restricted primal-dual gap* function. In order to derive from here convergence rates for the primal objective function, either Lipschitz continuity of g or finite dimensionality of the problem plus the condition that g must have full domain are necessary (see, for instance, [17, Theorem 9]). This means, that for infinite dimensional problems the assumptions required by both, PDHG and our method, for deriving convergence rates for the primal objective function are in fact equal, but for finite dimensional problems the assumption of PDHG are weaker. In either case, however, we are able to prove these rates for the sequence of iterates $(x_k)_{k\geq 1}$ itself whereas PDHG only has them for the sequence of so-called *ergodic iterates*, i.e. $({}^1\!/\!_k \sum_{i=1}^k x_i)_{k\geq 1}$, which is naturally undesirable as the averaging slows the convergence down. Furthermore, we do not show any convergence for the iterates as these are notoriously hard to obtain for accelerated method whereas PDHG gets these via standard fixed point arguments (see e.g. [105]).

Furthermore, we will also consider the case where only a stochastic oracle of the proximal operator of g is available to us. This setup corresponds e.g. to the case where the objective function is given as

$$\min_{x\in\mathbb{R}^d} f(x) + \sum_{i=1}^m g_i(A_i x), \tag{3.2}$$

where, for $i = 1, \dots, m$, $\mathbb{R}^n_i$ are real Hilbert spaces, $g_i : \mathbb{R}^n_i \to \mathbb{R}$ are convex and Lipschitz continuous functions and $A_i : \mathbb{R}^d \to \mathbb{R}^n_i$ are linear continuous operators, but the number of summands being large we wish to not compute all proximal operators of all $g_i, i = 1, \dots, m$, for purpose of making iterations cheaper to compute.

For the finite sum case (3.2), there exist algorithms of similar spirit such as those in [27, 83]. Some algorithms do in fact deal with a similar setup of stochastic gradient like evaluations, see [92], but only for smooth terms in the objective function.

In Section 3.2 we will cover useful identities and estimates connected to the Moreau envelope. In Section 3.3 we will deal with the deterministic case and prove a convergence rate of $\mathcal{O}({}^1\!/\!_k)$ for the function values at the iterates. Next up, in Section 3.4, we will consider the stochastic case as described above and prove a convergence rate of $\tilde{\mathcal{O}}({}^1\!/\!\sqrt{k})$. Last but not least, we will look at some numerical examples in image processing in Section 3.5.

It is important to note that the proof for the deterministic setting differs surprisingly from the one for the stochastic setting. The technique for the stochastic setting is less refined in the sense that there is no coupling between the smoothing parameter and the extrapolation parameter. Whereas this technique also works for the deterministic setting it gives a worse convergence rate of $\tilde{\mathcal{O}}({}^1\!/\!_k)$. The tight coupling of the two sequences of parameters, however, is not compatible with the particular stepsize requirements of the stochastic setting.

3.2 More properties of the Moreau envelope

As mentioned in the introduction, we want to smooth the nonsmooth summand of the objective function which is composed with the linear operator as this can be considered the crux of problem (3.1). The function $g \circ A$ will be *smoothed* via considering instead $g_\lambda \circ A : \mathbb{R}^d \to \mathbb{R}$. Clearly, by the chain rule, this function is continuously differentiable with gradient given for every $x \in \mathbb{R}^d$ by

$$\nabla \left(g_\lambda \circ A\right)(x) = A^* \nabla g_\lambda(Ax) = \frac{1}{\lambda} A^* \left(Ax - \operatorname{prox}_{\lambda g}(Ax)\right) = A^* \operatorname{prox}_{\frac{1}{\lambda} g^*}\left(\frac{Ax}{\lambda}\right),$$

where we used Lemma 2.3.7 to deduce the second and third equality. The gradient of $g_\lambda \circ A$ is thus Lipschitz continuous with Lipschitz constant $\frac{\|A\|^2}{\lambda}$, where $\|A\|$ denotes the operator norm of A.

The following lemmata have been presented in [101] in the finite dimensional case. We provide proofs in order to ensure that the statements hold true even in Hilbert spaces.

Lemma 3.2.1 (see [101, Lemma 10 (a)]). *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex and lower semicontinuous function. The maximizing argument in the definition of the Moreau envelope is given by its gradient, i.e. for $\lambda > 0$ it holds that*

$$\operatorname*{arg\,max}_{p \in \mathbb{R}^d} \left\{ \langle \cdot, p \rangle - g^*(p) - \frac{\lambda}{2} \|p\|^2 \right\} = \nabla g_\lambda(\cdot).$$

Proof. Let $x \in \mathbb{R}^d$ be fixed. It holds

$$\begin{aligned}
\operatorname*{arg\,max}_{p \in \mathbb{R}^d} \left\{ \langle x, p \rangle - g^*(p) - \frac{\lambda}{2} \|p\|^2 \right\} &= \operatorname*{arg\,max}_{p \in \mathbb{R}^d} \left\{ -\frac{1}{2\lambda} \|x\|^2 + \langle x, p \rangle - \frac{\lambda}{2} \|p\|^2 - g^*(p) \right\} \\
&= \operatorname*{arg\,max}_{p \in \mathbb{R}^d} \left\{ -\frac{\lambda}{2} \left\| \frac{x}{\lambda} - p \right\|^2 - g^*(p) \right\} \\
&= \operatorname*{arg\,min}_{p \in \mathbb{R}^d} \left\{ g^*(p) + \frac{\lambda}{2} \left\| \frac{x}{\lambda} - p \right\|^2 \right\} \\
&= \operatorname{prox}_{\frac{1}{\lambda} g^*}\left(\frac{x}{\lambda}\right)
\end{aligned}$$

and the conclusion follows by using Lemma 2.3.7. □

Lemma 3.2.2 (see [101, Lemma 10 (a)]). *For a proper, convex and lower semicontinuous function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and every $x \in \mathbb{R}^d$ we can consider the mapping from $(0, +\infty)$ to $\mathbb{R}$ given by*

$$\lambda \mapsto g_\lambda(x). \tag{3.3}$$

This mapping is convex and differentiable and its derivative is given by

$$\frac{\partial}{\partial \lambda} g_\lambda(x) = -\frac{1}{2} \|\nabla g_\lambda(x)\|^2 \qquad \forall x \in \mathbb{R}^d \ \forall \lambda \in (0, +\infty).$$

Proof. Let $x \in \mathbb{R}^d$ be fixed. From the definition of the Moreau envelope we can see that the mapping given in (3.3) is a pointwise supremum of a family of functions which are linear in λ. It is therefore convex. Furthermore, since the objective function is strongly concave, this supremum is uniquely attained at $\nabla g_\lambda(x) = \arg\max_{p\in\mathbb{R}^d} \left\{\langle x, p\rangle - g^*(p) - \frac{\lambda}{2}\|p\|^2\right\}$. According to the Danskin Theorem, the function $\lambda \mapsto g_\lambda(x)$ is differentiable and its derivative is given by

$$\begin{aligned}\frac{\partial}{\partial\lambda} g_\lambda(x) =& \frac{\partial}{\partial\lambda} \sup_{p\in\mathbb{R}^d} \left\{\langle x, p\rangle - g^*(p) - \frac{\lambda}{2}\|p\|^2\right\} \\ =& -\frac{1}{2}\|\nabla g_\lambda(x)\|^2 \quad \forall\lambda \in (0, +\infty).\end{aligned}$$

□

Lemma 3.2.3 (see [101, Lemma 10 (b)]). *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper, convex and lower semicontinuous. For $\lambda_1, \lambda_2 > 0$ and every $x \in \mathbb{R}^d$ it holds*

$$g_{\lambda_1}(x) \le g_{\lambda_2}(x) + (\lambda_2 - \lambda_1)\frac{1}{2}\|\nabla g_{\lambda_1}(x)\|^2. \tag{3.4}$$

If g is additionally L_g-Lipschitz and if $\lambda_2 \ge \lambda_1$, then

$$g_{\lambda_2}(x) \le g_{\lambda_1}(x) \le g_{\lambda_2}(x) + (\lambda_2 - \lambda_1)\frac{L_g^2}{2}. \tag{3.5}$$

Proof. Let $x \in \mathbb{R}^d$ be fixed. Via Lemma 3.2.2 we know that the map $\lambda \mapsto g_\lambda(x)$ is convex and differentiable. We can therefore use the gradient inequality to deduce that

$$\begin{aligned}g_{\lambda_2}(x) \ge& g_{\lambda_1}(x) + (\lambda_2 - \lambda_1)\left(\frac{\partial}{\partial\lambda} g_\lambda(x)\Big|_{\lambda=\lambda_1}\right) \\ =& g_{\lambda_1}(x) - (\lambda_2 - \lambda_1)\frac{1}{2}\|\nabla g_{\lambda_1}(x)\|^2,\end{aligned}$$

which is exactly the first statement of the lemma. The first inequality of (3.5) is obtained directly from the definition of the Moreau envelope whereas the second one follows from (3.4) together with Lemma 2.3.9. □

By applying a limiting argument it is easy to see that (3.5) implies that for any $\lambda > 0$

$$g_\lambda(x) \le g(x) \le g_\lambda(x) + \lambda\frac{L_g^2}{2} \tag{3.6}$$

which shows that the Moreau envelope is always a lower approximation of the original function.

Lemma 3.2.4 ([101, Lemma 10 (c)]). *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper, convex and lower semicontinuous. Then, for $\lambda > 0$ and every $x, y \in \mathbb{R}^d$ we have that*

$$g_\lambda(x) + \langle\nabla g_\lambda(x), y - x\rangle \le g(y) - \frac{\lambda}{2}\|\nabla g_\lambda(x)\|^2.$$

Proof. Using Lemma 3.2.1 and the definition of the Moreau envelope we get that

$$\begin{aligned}
g_\lambda(x) + \langle \nabla g_\lambda(x), y - x\rangle &= \langle x, \nabla g_\lambda(x)\rangle - g^*(\nabla g_\lambda(x)) - \frac{\lambda}{2}\|\nabla g_\lambda(x)\|^2 + \langle \nabla g_\lambda(x), y - x\rangle \\
&= \langle \nabla g_\lambda(x), y\rangle - g^*(\nabla g_\lambda(x)) - \frac{\lambda}{2}\|\nabla g_\lambda(x)\|^2 \\
&\leq \sup_{p\in\mathbb{R}^d}\{\langle p, y\rangle - g^*(p)\} - \frac{\lambda}{2}\|\nabla g_\lambda(x)\|^2 \\
&= g(y) - \frac{\lambda}{2}\|\nabla g_\lambda(x)\|^2.
\end{aligned}$$

□

In the convergence proof of Section 3.3 we will need the inequality in the above lemma at the points Ax and Ay, namely

$$\begin{aligned}
g(Ay) - \frac{\lambda}{2}\|\nabla g_\lambda(Ax)\|^2 &\geq g_\lambda(Ax) + \langle \nabla g_\lambda(Ax), Ay - Ax\rangle \\
&= g_\lambda(Ax) + \langle A^*\nabla g_\lambda(Ax), y - x\rangle \\
&= g_\lambda(Ax) + \langle \nabla(g_\lambda \circ A)(x), y - x\rangle \quad \forall x, y \in \mathbb{R}^d.
\end{aligned} \tag{3.7}$$

By applying Lemma 2.2.2 with g_λ, Ax and Ay instead of h, x and y respectively, we obtain

$$g_\lambda(Ax) + \langle \nabla(g_\lambda \circ A)(x), y - x\rangle \leq g_\lambda(Ay) - \frac{\lambda}{2}\|\nabla g_\lambda(Ax) - \nabla g_\lambda(Ay)\|^2 \quad \forall x, y \in \mathbb{R}^d. \tag{3.8}$$

3.3 Deterministic method

The idea of the algorithm which we propose to solve (3.1) is to smooth g and then to solve the resulting problem by means of an accelerated proximal gradient method.

Algorithm 3.3.1 (Variable Accelerated SmooThing (VAST))**.** Let $y_0 = x_0 \in \mathbb{R}^d$, $(\lambda_k)_{k\geq 0} \subseteq (0, +\infty)$, and $(t_k)_{k\geq 1}$ a sequence of real numbers with $t_1 = 1$ and $t_k \geq 1$ for every $k \geq 2$. Consider the following iterative scheme

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l}
L_k = \frac{\|A\|^2}{\lambda_k} \\
\gamma_k = \frac{1}{L_k} \\
x_k = \operatorname{prox}_{\gamma_k f}\left(y_{k-1} - \gamma_k A^* \operatorname{prox}_{\frac{1}{\lambda_k} g^*}\left(A\frac{y_{k-1}}{\lambda_k}\right)\right) \\
y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).
\end{array}\right.$$

Remark 3.3.1. The assumption $t_1 = 1$ can be removed but guarantees easier computation and is also in line with classical choices of $(t_k)_{k\geq 1}$ in [26, 73].

Remark 3.3.2. The sequence $(u_k)_{k\geq 1}$ given by

$$u_k := x_{k-1} + t_k(x_k - x_{k-1}) \quad \forall k \geq 1,$$

despite not appearing in the algorithm, will feature a prominent role in the convergence proof. Due to the convention $t_1 = 1$ we have that

$$u_1 := x_0 + t_1(x_1 - x_0) = x_1.$$

We also denote

$$F^k = f + g_{\lambda_k} \circ A \quad \forall k \geq 0.$$

The next theorem is the main result of this section and it will play a fundamental role when proving a convergence rate of $\mathcal{O}(1/k)$ for the sequence $(F(x_k))_{k\geq 0}$.

Theorem 3.3.2. *Consider the setup of* (3.1) *and let* $(x_k)_{k\geq 0}$ *and* $(y_k)_{k\geq 0}$ *be the sequences generated by Algorithm 3.3.1. Assume that for every* $k \geq 1$

$$\lambda_k - \lambda_{k+1} - \frac{\lambda_{k+1}}{t_{k+1}} \leq 0$$

and

$$\left(1 - \frac{1}{t_{k+1}}\right)\gamma_{k+1}t_{k+1}^2 = \gamma_k t_k^2.$$

Then, for every optimal solution x^* *of* (3.1), *it holds*

$$F(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \lambda_N \frac{L_g^2}{2} \quad \forall N \geq 1.$$

The proof of this result relies on several partial results which we will prove as follows.

Lemma 3.3.3. *The following statement holds for every* $z \in \mathbb{R}^d$ *and every* $k \geq 0$

$$F^{k+1}(x_{k+1}) + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq$$
$$f(z) + g_{\lambda_{k+1}}(Ay_k) + \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), z - y_k \right\rangle + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2.$$

Proof. Let $k \geq 0$ be fixed. Since, by the definition of the proximal map, x_{k+1} is the minimizer of a $\frac{1}{\gamma_{k+1}}$-strongly convex function we deduce from Lemma 2.1.11 that for every $z \in \mathbb{R}^d$

$$f(x_{k+1}) + g_{\lambda_{k+1}}(Ay_k) + \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), x_{k+1} - y_k \right\rangle + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - y_k\|^2 +$$
$$\frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq f(z) + g_{\lambda_{k+1}}(Ay_k) + \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), z - y_k \right\rangle + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2.$$

Next we use the L_{k+1}-smoothness of $g_{\lambda_{k+1}} \circ A$ and the fact that $\frac{1}{\gamma_{k+1}} = L_{k+1}$ to deduce that

$$\begin{aligned} f(x_{k+1}) + g_{\lambda_{k+1}}(Ax_{k+1}) + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq& \\ & f(z) + g_{\lambda_{k+1}}(Ay_k) + \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), z - y_k \right\rangle + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2, \end{aligned}$$

which finishes the proof. □

Lemma 3.3.4. *Let x^* be an optimal solution of* (3.1). *Then it holds*

$$\gamma_1(F^1(x_1) - F(x^*)) + \frac{1}{2}\|u_1 - x^*\|^2 \leq \frac{1}{2}\|x^* - x_0\|^2.$$

Proof. We use the gradient inequality to deduce that for every $z \in \mathbb{R}^d$ and every $k \geq 0$

$$g_{\lambda_{k+1}}(Ay_k) + \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), z - y_k \right\rangle \leq g_{\lambda_{k+1}}(Az) \leq g(Az)$$

and plug this into the statement of Lemma 3.3.3 to conclude that

$$F^{k+1}(x_{k+1}) + \frac{1}{2\gamma_{k+1}}\|x_{k+1} - z\|^2 \leq F(z) + \frac{1}{2\gamma_{k+1}}\|z - y_k\|^2.$$

For $k = 0$ we get that

$$F^1(x_1) + \frac{1}{2\gamma_1}\|x_1 - x^*\|^2 \leq F(x^*) + \frac{1}{2\gamma_1}\|x^* - y_0\|^2.$$

Now we us the fact that $u_1 = x_1$ and $y_0 = x_0$ to obtain the conclusion. □

Lemma 3.3.5. *Let x^* be an optimal solution of* (3.1). *The following descent-type inequality holds for every $k \geq 0$*

$$\begin{aligned} F^{k+1}(x_{k+1}) - F(x^*) + \frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \leq& \left(1 - \frac{1}{t_{k+1}}\right)\left(F^k(x_k) - F(x^*)\right) + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \\ &+ \left(1 - \frac{1}{t_{k+1}}\right)\left(\lambda_k - \lambda_{k+1} - \frac{\lambda_{k+1}}{t_{k+1}}\right)\|\nabla g_{\lambda_{k+1}}(Ax_k)\|^2. \end{aligned}$$

Proof. Let $k \geq 0$ be fixed. We apply Lemma 3.3.3 with $z := \left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*$ to deduce that

$$\begin{aligned} &F^{k+1}(x_{k+1}) + \frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} \\ &\quad\leq f\left(\left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*\right) + g_{\lambda_{k+1}}(Ay_k) + \frac{1}{t_{k+1}}\left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), x^* - y_k \right\rangle \\ &\qquad+ \left(1 - \frac{1}{t_{k+1}}\right)\left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), x_k - y_k \right\rangle + \frac{1}{2\gamma_{k+1}t_{k+1}^2}\|u_k - x^*\|^2. \end{aligned}$$

Using the convexity of f gives

$$\begin{aligned}
&F^{k+1}(x_{k+1}) + \frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1} t_{k+1}^2} \\
&\quad \leq \left(1 - \frac{1}{t_{k+1}}\right) f(x_k) + \frac{1}{t_{k+1}} f(x^*) + \left(1 - \frac{1}{t_{k+1}}\right) g_{\lambda_{k+1}}(Ay_k) \\
&\qquad + \left(1 - \frac{1}{t_{k+1}}\right) \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), x_k - y_k \right\rangle \\
&\qquad + \frac{1}{t_{k+1}} g_{\lambda_{k+1}}(Ay_k) + \frac{1}{t_{k+1}} \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), x^* - y_k \right\rangle + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1} t_{k+1}^2}.
\end{aligned} \tag{3.9}$$

Now, we use (3.7) to deduce that

$$g_{\lambda_{k+1}}(Ay_k) + \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), x^* - y_k \right\rangle \leq g(Ax^*) - \frac{\lambda_{k+1}}{2} \|\nabla g_{\lambda_{k+1}}(Ay_k)\|^2 \tag{3.10}$$

and (3.8) to conclude that

$$\begin{aligned}
&g_{\lambda_{k+1}}(Ay_k) + \left\langle \nabla(g_{\lambda_{k+1}} \circ A)(y_k), x_k - y_k \right\rangle \\
&\quad \leq g_{\lambda_{k+1}}(Ax_k) - \frac{\lambda_{k+1}}{2} \|\nabla g_{\lambda_{k+1}}(Ay_k) - \nabla g_{\lambda_{k+1}}(Ax_k)\|^2.
\end{aligned} \tag{3.11}$$

Combining (3.9), (3.10) and (3.11) gives

$$\begin{aligned}
&F^{k+1}(x_{k+1}) + \frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1} t_{k+1}^2} \\
&\quad \leq \left(1 - \frac{1}{t_{k+1}}\right) g_{\lambda_{k+1}}(Ax_k) + \left(1 - \frac{1}{t_{k+1}}\right) f(x_k) + \frac{1}{t_{k+1}} g(Ax^*) + \frac{1}{t_{k+1}} f(x^*) \\
&\qquad - \left(1 - \frac{1}{t_{k+1}}\right) \frac{\lambda_{k+1}}{2} \|\nabla g_{\lambda_{k+1}}(Ay_k) - \nabla g_{\lambda_{k+1}}(Ax_k)\|^2 \\
&\qquad - \frac{1}{t_{k+1}} \frac{\lambda_{k+1}}{2} \|\nabla g_{\lambda_{k+1}}(Ay_k)\|^2 + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1} t_{k+1}^2}.
\end{aligned}$$

The first term on the right hand side is $g_{\lambda_{k+1}}(Ax_k)$ but we would like it to be $g_{\lambda_k}(Ax_k)$. Therefore we use Lemma 3.2.3 to deduce that

$$\begin{aligned}
&F^{k+1}(x_{k+1}) + \frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1} t_{k+1}^2} \\
&\quad \leq \left(1 - \frac{1}{t_{k+1}}\right) F^k(x_k) + \frac{1}{t_{k+1}} F(x^*) \\
&\qquad + \left(1 - \frac{1}{t_{k+1}}\right) (\lambda_k - \lambda_{k+1}) \frac{1}{2} \|\nabla g_{\lambda_{k+1}}(Ax_k)\|^2 + \frac{\|u_k - x^*\|^2}{2\gamma_{k+1} t_{k+1}^2} \\
&\qquad - \left(1 - \frac{1}{t_{k+1}}\right) \frac{\lambda_{k+1}}{2} \|\nabla g_{\lambda_{k+1}}(Ay_k) - \nabla g_{\lambda_{k+1}}(Ax_k)\|^2 - \frac{\lambda_{k+1}}{2t_{k+1}} \|\nabla g_{\lambda_{k+1}}(Ay_k)\|^2.
\end{aligned} \tag{3.12}$$

Next up we want to estimate all the norms of gradients by using Lemma 2.1.2 which says that

$$\begin{aligned}\left(1-\frac{1}{t_{k+1}}\right)\|\nabla g_{\lambda_{k+1}}(Ay_k)-\nabla g_{\lambda_{k+1}}(Ax_k)\|^2+&\frac{1}{t_{k+1}}\|\nabla g_{\lambda_{k+1}}(Ay_k)\|^2\\ \geq\left(1-\frac{1}{t_{k+1}}\right)\frac{1}{t_{k+1}}\|\nabla g_{\lambda_{k+1}}(Ax_k)\|^2.&\end{aligned}\tag{3.13}$$

Combining (3.12) and (3.13) gives

$$\begin{aligned}&F^{k+1}(x_{k+1})+\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}\\ &\quad\leq\left(1-\frac{1}{t_{k+1}}\right)F^k(x_k)+\frac{1}{t_{k+1}}F(x^*)+\left(1-\frac{1}{t_{k+1}}\right)(\lambda_k-\lambda_{k+1})\frac{1}{2}\|\nabla g_{\lambda_{k+1}}(Ax_k)\|^2\\ &\quad\quad-\frac{\lambda_{k+1}}{2}\left(1-\frac{1}{t_{k+1}}\right)\frac{1}{t_{k+1}}\|\nabla g_{\lambda_{k+1}}(Ax_k)\|^2+\frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}.\end{aligned}$$

Now we combine the two terms containing $\|\nabla g_{\lambda_{k+1}}(Ax_k)\|^2$ and get that

$$\begin{aligned}&F^{k+1}(x_{k+1})+\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}\\ &\quad\leq\left(1-\frac{1}{t_{k+1}}\right)F^k(x_k)+\frac{1}{t_{k+1}}F(x^*)+\frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}\\ &\quad\quad+\left(1-\frac{1}{t_{k+1}}\right)\left(\lambda_k-\lambda_{k+1}-\frac{\lambda_{k+1}}{t_{k+1}}\right)\frac{1}{2}\|\nabla g_{\lambda_{k+1}}(Ax_k)\|^2.\end{aligned}$$

By subtracting $F(x^*)=f(x^*)+g(Ax^*)$ on both sides we obtain the desired statement. □

Now we are in the position to prove Theorem 3.3.2.

Proof of Theorem 3.3.2. We start with the statement of Lemma 3.3.5 and use the assumption that

$$\lambda_k-\lambda_{k+1}-\frac{\lambda_{k+1}}{t_{k+1}}\leq 0$$

to make the last term in the inequality disappear for every $k\geq 0$

$$F^{k+1}(x_{k+1})-F(x^*)+\frac{\|u_{k+1}-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}\leq\left(1-\frac{1}{t_{k+1}}\right)\Big(F^k(x_k)-F(x^*)\Big)+\frac{\|u_k-x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}.$$

Now we use the assumption that

$$\left(1-\frac{1}{t_{k+1}}\right)\gamma_{k+1}t_{k+1}^2=\gamma_k t_k^2$$

to get that for every $k \geq 0$

$$\gamma_{k+1} t_{k+1}^2 (F^{k+1}(x_{k+1}) - F(x^*)) + \frac{\|u_{k+1} - x^*\|^2}{2} \leq \gamma_k t_k^2 (F^k(x_k) - F(x^*)) + \frac{\|u_k - x^*\|^2}{2}. \tag{3.14}$$

Let $N \geq 2$. Summing (3.14) from $k = 1$ to $N - 1$ and getting rid of the nonnegative term $\|u_N - x^*\|^2$ gives

$$\gamma_N t_N^2 (F^N(x_N) - F(x^*)) \leq \gamma_1 (F^1(x_1) - F(x^*)) + \frac{\|u_1 - x^*\|^2}{2} \quad \forall N \geq 2.$$

Since $t_1 = 1$, the above inequality is fulfilled also for $N = 1$. Using Lemma 3.3.4 shows that

$$F^N(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{\gamma_N t_N^2} \quad \forall N \geq 1.$$

The above inequality, however, is still in terms of the smoothed objective function. In order to go to the actual objective function we apply (3.6) and deduce that

$$F(x_N) - F(x^*) \leq F^N(x_N) - F(x^*) + \lambda_N \frac{L_g^2}{2} \leq \frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \lambda_N \frac{L_g^2}{2} \quad \forall N \geq 1.$$

□

Corollary 3.3.6. *By choosing the parameters $(\lambda_k)_{k\geq 1}, (t_k)_{k\geq 1}, (\gamma_k)_{k\geq 1}$ in the following way,*

$$t_1 = 1, \quad \lambda_1 = b\|A\|^2, \text{ for } b > 0,$$

and for every $k \geq 1$

$$t_{k+1} := \sqrt{t_k^2 + 2t_k}, \quad \lambda_{k+1} := \lambda_k \frac{t_k^2}{t_{k+1}^2 - t_{k+1}}, \quad \gamma_k := \frac{\lambda_k}{\|A\|^2}, \tag{3.15}$$

they fulfill

$$\lambda_k - \lambda_{k+1} - \frac{\lambda_{k+1}}{t_{k+1}} \leq 0 \tag{3.16}$$

and

$$\left(1 - \frac{1}{t_{k+1}}\right) \gamma_{k+1} t_{k+1}^2 = \gamma_k t_k^2 \tag{3.17}$$

For this choice of the parameters we have that

$$F(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{b(N+1)} + \frac{bL_g^2\|A\|^2}{(N+1)} \exp\left(\frac{4\pi^2}{6}\right) \quad \forall N \geq 1.$$

Proof. Since γ_k and λ_k are a scalar multiple of each other, (3.17) is equivalent to

$$\left(1 - \frac{1}{t_{k+1}}\right) \lambda_{k+1} t_{k+1}^2 = \lambda_k t_k^2 \quad \forall k \geq 1$$

and further to (by taking into account that $t_{k+1} > 1$ for every $k \geq 1$)

$$\lambda_{k+1} = \lambda_k \frac{t_k^2}{t_{k+1}^2} \frac{t_{k+1}}{t_{k+1}-1} = \lambda_k \frac{t_k^2}{t_{k+1}^2 - t_{k+1}} \quad \forall k \geq 1. \tag{3.18}$$

Our update scheme in (3.15) for the sequence $(\lambda_k)_{k\geq 1}$ is exactly chosen in such a way that it satisfies this. Plugging (3.18) into (3.16) gives for every $k \geq 1$ the condition

$$1 \leq \left(1 + \frac{1}{t_{k+1}}\right) \frac{t_k^2}{t_{k+1}^2} \frac{t_{k+1}}{t_{k+1}-1} = \frac{t_k^2}{t_{k+1}^2} \frac{t_{k+1}+1}{t_{k+1}-1},$$

which is equivalent to

$$0 \geq t_{k+1}^3 - t_{k+1}^2 - t_k^2 t_{k+1} - t_k^2$$

and further to

$$t_{k+1}^2 + t_k^2 \geq t_{k+1}\left(t_{k+1}^2 - t_k^2\right).$$

Plugging in $t_{k+1} = \sqrt{t_k^2 + 2t_k}$ we get that this equivalent to

$$t_{k+1}^2 + t_k^2 \geq t_{k+1} 2t_k \quad \forall k \geq 1,$$

which is evidently fulfilled. Thus, the choices in (3.15) are indeed feasible for our algorithm.

Now we want to prove the claimed convergence rates. Via induction we show that

$$\frac{k+1}{2} \leq t_k \leq k \quad \forall k \geq 1. \tag{3.19}$$

Evidently, this holds for $t_1 = 1$. Assuming that (3.19) holds for $k \geq 1$, we easily see that

$$t_{k+1} = \sqrt{t_k^2 + 2t_k} \leq \sqrt{k^2 + 2k} \leq \sqrt{k^2 + 2k + 1} = k + 1$$

and, on the other hand,

$$t_{k+1} = \sqrt{t_k^2 + 2t_k} \geq \sqrt{\frac{(k+1)^2}{4} + k + 1} = \frac{1}{2}\sqrt{k^2 + 6k + 5} \geq \frac{1}{2}\sqrt{k^2 + 4k + 4} = \frac{k+2}{2}.$$

In the following we prove a similar estimate for the sequence $(\lambda_k)_{k\geq 1}$. To this end we show, again by induction, the following recursion for every $k \geq 2$

$$\lambda_k = \lambda_1 \frac{\prod_{j=1}^{k-1} t_j}{\prod_{j=2}^{k} (t_j - 1)} \frac{1}{t_k}. \tag{3.20}$$

For $k = 2$ this follows from the definition (3.18). Assume now that (3.20) holds for $k \geq 2$. From here we have that

$$\lambda_{k+1} = \lambda_k \frac{t_k^2}{t_{k+1}(t_{k+1}-1)} = \lambda_1 \frac{\prod_{j=1}^{k-1} t_j}{\prod_{j=2}^{k} (t_j - 1)} \frac{1}{t_k} \frac{t_k^2}{t_{k+1}(t_{k+1}-1)} = \lambda_1 \frac{\prod_{j=1}^{k} t_j}{\prod_{j=2}^{k+1} (t_j - 1)} \frac{1}{t_{k+1}}.$$

Using (3.20) together with (3.19) we can check that for every $k \geq 1$

$$\lambda_{k+1} = \lambda_1 \frac{\prod_{j=1}^{k} t_j}{\prod_{j=2}^{k+1}(t_j - 1)} \frac{1}{t_{k+1}} = \frac{\lambda_1}{t_{k+1}} \prod_{j=1}^{k} \frac{t_j}{(t_{j+1} - 1)} \geq \frac{\lambda_1}{t_{k+1}} = b\|A\|^2 \frac{1}{t_{k+1}}, \tag{3.21}$$

where we used in the last step the fact that $t_{k+1} \leq t_k + 1$.

The last thing to check is the fact that λ_k goes to zero like $\frac{1}{k}$. First we check that for every $k \geq 1$

$$\frac{t_k}{t_{k+1} - 1} \leq 1 + \frac{1}{t_{k+1}(t_{k+1} - 1)}. \tag{3.22}$$

This can be seen via

$$(t_k + 1)t_{k+1} \leq (t_k + 1)^2 = t_{k+1}^2 + 1 \quad \forall k \geq 1.$$

By bringing t_{k+1} to the other side we get that

$$t_{k+1}t_k \leq t_{k+1}^2 - t_{k+1} + 1,$$

from which we can deduce (3.22) by dividing by $t_{k+1}^2 - t_{k+1}$.

We plug in the estimate (3.22) in (3.20) and get for every $k \geq 2$

$$\begin{aligned}
\lambda_k &= \lambda_1 \frac{\prod_{j=1}^{k-1} t_j}{\prod_{j=1}^{k-1}(t_{j+1} - 1)} \frac{1}{t_k} \\
&\leq \lambda_1 \prod_{j=1}^{k-1} \left(1 + \frac{1}{t_{j+1}(t_{j+1} - 1)}\right) \frac{1}{t_k} \leq \lambda_1 \prod_{j=1}^{k-1} \left(1 + \frac{4}{(j+2)j}\right) \frac{1}{t_k} \\
&\leq \lambda_1 \prod_{j=1}^{k-1} \left(1 + \frac{4}{j^2}\right) \frac{1}{t_k} \leq \lambda_1 \exp\left(\frac{\pi^2 4}{6}\right) \frac{1}{t_k} = b\|A\|^2 \exp\left(\frac{\pi^2 4}{6}\right) \frac{1}{t_k}.
\end{aligned}$$

With the above inequalities we can to deduce the claimed convergence rates. First note that from Theorem 3.3.2 we have

$$F(x_N) - F(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \lambda_N \frac{L_g^2}{2} \quad \forall N \geq 1.$$

Now, in order to obtain the desired conclusion, we use the above estimates and deduce for every $N \geq 1$

$$\begin{aligned}
\frac{\|x_0 - x^*\|^2}{2\gamma_N t_N^2} + \lambda_N \frac{L_g^2}{2} &\leq \frac{\|x_0 - x^*\|^2}{2bt_N} + \frac{bL_g^2\|A\|^2}{2t_N} \exp\left(\frac{4\pi^2}{6}\right) \\
&\leq \frac{\|x_0 - x^*\|^2}{b(N+1)} + \frac{bL_g^2\|A\|^2}{(N+1)} \exp\left(\frac{4\pi^2}{6}\right),
\end{aligned}$$

where we used that

$$\gamma_N t_N = \frac{\lambda_N t_N}{\|A\|^2} \geq b,$$

as shown in (3.21). □

Remark 3.3.3. Consider the choice (see [73])

$$t_1 = 1, \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \quad \forall k \geq 1$$

and

$$\lambda_1 = b\|A\|^2, \text{ for } b > 0.$$

Since

$$t_k^2 = t_{k+1}^2 - t_{k+1} \quad \forall k \geq 1,$$

we see that in this setting we have to choose

$$\lambda_k = b\|A\|^2 \text{ and } \gamma_k = b \quad \forall k \geq 1.$$

Thus, the sequence of optimal function values $(F(x_N))_{N\geq 1}$ approaches a $b\|A\|^2 \frac{L_g}{2}$-approximation of the optimal objective value $F(x^*)$ with a convergence rate of $\mathcal{O}(\frac{1}{N^2})$, i.e.

$$F(x_N) - F(x^*) \leq 2\frac{\|x_0 - x^*\|^2}{b(N+1)^2} + b\frac{\|A\|^2 L_g^2}{2} \quad \forall N \geq 1.$$

3.4 Stochastic method

The problem is the same as in the deterministic case other than the fact that at each iteration we are only given a *stochastic estimator* of the quantity

$$\nabla(g_{\lambda_k} \circ A)(\cdot) = A^* \operatorname{prox}_{\frac{1}{\lambda_k} g^*}\left(\frac{1}{\lambda_k} A(\cdot)\right) \quad \forall k \geq 1.$$

Remark 3.4.1. Consider Algorithm 3.4.7 for a setting where such an estimator is easily computed.

For the stochastic quantities arising in this section we will use the following notation. For every $k \geq 0$, we denote by $\sigma(x_0, \ldots, x_k)$ the smallest σ-algebra generated by the family of random variables $\{x_0, \ldots, x_k\}$ and by $\mathbb{E}_k(\cdot) := \mathbb{E}[\cdot \mid \sigma(x_0, \ldots, x_k)]$ the conditional expectation with respect to this σ-algebra.

Algorithm 3.4.1 (stochastic Variable Accelerated SmooThing (sVAST))**.** Let $y_0 = x_0 \in \mathbb{R}^d$, $(\lambda_k)_{k\geq 1}$ a sequence of positive and nonincreasing real numbers, and $(t_k)_{k\geq 1}$ a sequence of real numbers with $t_1 = 1$ and $t_k \geq 1$ for every $k \geq 2$. Consider the following iterative scheme

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} L_k = \frac{\|A\|^2}{\lambda_k} \\ \gamma_k = \frac{1}{L_k} \\ x_k = \operatorname{prox}_{\gamma_k f}(y_{k-1} - \gamma_k \xi_{k-1}) \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}), \end{array} \right.$$

where we make the standard assumptions about our gradient estimator of being unbiased, i.e.

$$\mathbb{E}_k[\xi_k] = \nabla(g_{\lambda_{k+1}} \circ A)(y_k),$$

and having bounded variance

$$\mathbb{E}_k \left[\|\xi_k - \nabla(g_{\lambda_{k+1}} \circ A)(y_k)\|^2\right] \leq \sigma^2$$

for every $k \geq 0$.

Note that we use the same notations as in the deterministic case

$$u_k := x_{k-1} + t_k(x_k - x_{k-1}) \text{ and } F^k(\cdot) := f + g_{\lambda_k} \circ A \quad \forall k \geq 1.$$

Lemma 3.4.2. *The following statement holds for every (deterministic) $z \in \mathbb{R}^d$ and every $k \geq 0$*

$$\mathbb{E}_k \left[F^{k+1}(x_{k+1}) + \frac{\|x_{k+1} - z\|^2}{2\gamma_{k+1}}\right] \leq \; F^{k+1}(z) + \frac{\|z - y_k\|^2}{2\gamma_{k+1}} + \gamma_{k+1}\left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right).$$

Proof. Here we have to proceed a little bit different from Lemma 3.3.3. Namely, we have to treat the gradient step and the proximal step differently. For this purpose we define the auxiliary variable

$$z_k := y_{k-1} - \gamma_k \xi_{k-1} \quad \forall k \geq 1.$$

Let $k \geq 1$ and $z \in \mathbb{R}^d$ be arbitrary but fixed. From the gradient step we get

$$\begin{aligned}\|z - z_k\|^2 &= \|y_{k-1} - \gamma_k \xi_{k-1} - z\|^2 \\ &= \|y_{k-1} - z\|^2 - 2\gamma_k \langle \xi_{k-1}, y_{k-1} - z\rangle + \gamma_k^2 \|\xi_{k-1}\|^2.\end{aligned}$$

Taking the conditional expectation gives

$$\mathbb{E}_{k-1}\left[\|z - z_k\|^2\right] = \|y_{k-1} - z\|^2 - 2\gamma_k \langle \nabla(g_{\lambda_k} \circ A)(y_{k-1}), y_{k-1} - z\rangle + \gamma_k^2 \mathbb{E}_{k-1}\left[\|\xi_{k-1}\|^2\right].$$

Using the gradient inequality we deduce

$$\begin{aligned}\mathbb{E}_{k-1}\left[\|z - z_k\|^2\right] \leq \|y_{k-1} - z\|^2 &- 2\gamma_k((g_{\lambda_k} \circ A)(y_{k-1}) - (g_{\lambda_k} \circ A)(z)) \\ &+ \gamma_k^2 \mathbb{E}_{k-1}\left[\|\xi_{k-1}\|^2\right]\end{aligned}$$

and therefore

$$\begin{aligned}&\gamma_k(g_{\lambda_k} \circ A)(y_{k-1}) + \frac{1}{2}\mathbb{E}_{k-1}\left[\|z - z_k\|^2\right] \\ &\quad \leq \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k(g_{\lambda_k} \circ A)(z) + \frac{\gamma_k^2}{2}\mathbb{E}_{k-1}\left[\|\xi_{k-1}\|^2\right].\end{aligned} \tag{3.23}$$

Also from the smoothness of $g_{\lambda_k} \circ A$ we deduce via the descent lemma that

$$g_{\lambda_k}(Az_k) \leq g_{\lambda_k}(Ay_{k-1}) + \langle \nabla(g_{\lambda_k} \circ A)(y_{k-1}), z_k - y_{k-1}\rangle + \frac{L_k}{2}\|z_k - y_{k-1}\|^2.$$

Plugging in the definition of z_k and using the fact that $L_k = \frac{1}{\gamma_k}$ we get

$$g_{\lambda_k}(Az_k) \leq g_{\lambda_k}(Ay_{k-1}) - \gamma_k \left\langle \nabla(g_{\lambda_k} \circ A)(y_{k-1}), \xi_{k-1} \right\rangle + \frac{\gamma_k}{2} \|\xi_{k-1}\|^2.$$

Now we take the conditional expectation to obtain that

$$\mathbb{E}_{k-1}[g_{\lambda_k}(Az_k)] \leq g_{\lambda_k}(Ay_{k-1}) - \gamma_k \left\| \nabla(g_{\lambda_k} \circ A)(y_{k-1}) \right\|^2 + \frac{\gamma_k}{2} \mathbb{E}_{k-1}\left[\|\xi_{k-1}\|^2\right]. \quad (3.24)$$

Multiplying (3.24) by γ_k and adding it to (3.23) gives

$$\begin{aligned}&\gamma_k \mathbb{E}_{k-1}\left[g_{\lambda_k}(Az_k)\right] + \frac{1}{2}\mathbb{E}_{k-1}\left[\|z - z_k\|^2\right] \leq \\ &\qquad \gamma_k g_{\lambda_k}(Az) + \frac{1}{2}\|y_{k-1} - z\|^2 - \gamma_k^2 \left\| \nabla(g_{\lambda_k} \circ A)(y_{k-1}) \right\|^2 + \gamma_k^2 \mathbb{E}_{k-1}\left[\|\xi_{k-1}\|^2\right].\end{aligned}$$

Now we use the assumption about the bounded variance to conclude that

$$\gamma_k \mathbb{E}_{k-1}\left[g_{\lambda_k}(Az_k)\right] + \frac{1}{2}\mathbb{E}_{k-1}\left[\|z - z_k\|^2\right] \leq \gamma_k g_{\lambda_k}(Az) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2 \sigma^2. \quad (3.25)$$

Next up for the proximal step we deduce

$$f(x_k) + \frac{1}{2\gamma_k}\|x_k - z_k\|^2 + \frac{1}{2\gamma_k}\|x_k - z\|^2 \leq f(z) + \frac{1}{2\gamma_k}\|z - z_k\|^2. \quad (3.26)$$

Taking the conditional expectation and combining (3.25) and (3.26) we get

$$\begin{aligned}&\mathbb{E}_{k-1}\left[\gamma_k(g_{\lambda_k}(Az_k) + f(x_k)) + \frac{1}{2}\|x_k - z_k\|^2 + \frac{1}{2}\|x_k - z\|^2\right] \\ &\qquad \leq \gamma_k F^k(z) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2 \sigma^2.\end{aligned}$$

From here, using now Lemma 2.3.9, we get that

$$\begin{aligned}&\mathbb{E}_{k-1}\left[\gamma_k F^k(x_k) - \gamma_k L_g \|A\| \|x_k - z_k\| + \frac{1}{2}\|x_k - z_k\|^2 + \frac{1}{2}\|x_k - z\|^2\right] \\ &\qquad \leq \gamma_k F^k(z) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2 \sigma^2.\end{aligned}$$

Now we use

$$-\frac{1}{2}\gamma_k^2 L_g^2 \|A\|^2 \leq \frac{1}{2}\|x_k - z_k\|^2 - \gamma_k L_g \|A\| \|x_k - z_k\|$$

to obtain that

$$\mathbb{E}_{k-1}\left[\gamma_k F^k(x_k) + \frac{1}{2}\|x_k - z\|^2\right] \leq \gamma_k F^k(z) + \frac{1}{2}\|y_{k-1} - z\|^2 + \gamma_k^2 \sigma^2 + \frac{1}{2}\gamma_k^2 L_g^2 \|A\|^2.$$

□

Lemma 3.4.3. *Let x^* be an optimal solution of 3.1. Then it holds*

$$\mathbb{E}\big[\gamma_1(F^1(x_1) - F^1(x^*))\big] + \frac{1}{2}\|u_1 - x^*\|^2 \leq \frac{1}{2}\|x_0 - x^*\|^2 + \gamma_1^2\sigma^2 + \frac{1}{2}\gamma_1^2 L_g^2\|A\|^2.$$

Proof. Applying the previous lemma with $k = 0$ and $z = x^*$, we get that

$$\mathbb{E}\left[\gamma_1 F^1(x_1) + \frac{1}{2}\|x_1 - x^*\|^2\right] \leq \gamma_1 F^1(x^*) + \frac{1}{2}\|y_0 - x^*\|^2 + \gamma_1^2\sigma^2 + \frac{1}{2}\gamma_1^2 L_g^2\|A\|^2.$$

Therefore, using the fact that $y_0 = x_0$ and $u_1 = x_1$,

$$\mathbb{E}\left[\gamma_1(F^1(x_1) - F^1(x^*)) + \frac{1}{2}\|u_1 - x^*\|^2\right] \leq \frac{1}{2}\|x_0 - x^*\|^2 + \gamma_1^2\sigma^2 + \frac{1}{2}\gamma_1^2 L_g^2\|A\|^2,$$

which finishes the proof. □

Theorem 3.4.4. *Consider the setup of 3.1 and let $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ denote the sequences generated by Algorithm 3.4.1. Assume that for all $k \geq 1$*

$$\rho_{k+1} := t_k^2 - t_{k+1}^2 + t_{k+1} \geq 0.$$

Then, for every optimal solution x^* of 3.1, it holds

$$\begin{aligned}\mathbb{E}[F(x_N) - F(x^*)] \leq{}& \frac{1}{\gamma_N t_N^2}\frac{1}{2}\|x_0 - x^*\|^2 + \frac{1}{\gamma_N t_N^2}\frac{\|A\|^2 L_g^2}{2}\sum_{k=1}^{N}\gamma_k^2(t_k + \rho_k)\\ &+ \frac{1}{\gamma_N t_N^2}\left(\sigma^2 + \frac{\|A\|^2 L_g^2}{2}\right)\sum_{k=1}^{N} t_k^2\gamma_k^2 \quad \forall N \geq 1.\end{aligned}$$

Proof of Theorem 3.4.4. Let $k \geq 0$ be fixed. Lemma 3.4.2 for $z := \left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*$ gives

$$\mathbb{E}_k\left[F^{k+1}(x_{k+1}) + \frac{1}{2\gamma_{k+1}}\left\|\frac{1}{t_{k+1}}u_{k+1} - \frac{1}{t_{k+1}}x^*\right\|^2\right] \leq$$
$$F^{k+1}\left(\left(1 - \frac{1}{t_{k+1}}\right)x_k + \frac{1}{t_{k+1}}x^*\right) + \frac{1}{2\gamma_{k+1}}\left\|\frac{1}{t_{k+1}}x^* - \frac{1}{t_{k+1}}u_k\right\|^2 + \gamma_{k+1}\left(\sigma^2 + \frac{\|A\|^2 L_g^2}{2}\right).$$

From here and from the convexity of F^{k+1} follows

$$\begin{aligned}\mathbb{E}_k\left[F^{k+1}(x_{k+1}) - F^{k+1}(x^*)\right] - \left(1 - \frac{1}{t_{k+1}}\right)(F^{k+1}(x_k) - F^{k+1}(x^*)) \leq{}&\\ \frac{\|u_k - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2} - \mathbb{E}_k\left[\frac{\|u_{k+1} - x^*\|^2}{2\gamma_{k+1}t_{k+1}^2}\right] + \gamma_{k+1}\left(\sigma^2 + \frac{\|A\|^2}{2}\right).&\end{aligned}$$

Let us now introduce for convenience the notation $\Delta_k := F^k(x_k) - F^k(x^*)$ for all $k \geq 0$. By multiplying both sides with by t_{k+1}^2, we deduce

$$\begin{aligned}
&\mathbb{E}_k\left[t_{k+1}^2 \Delta_{k+1}\right] + (t_{k+1} - t_{k+1}^2)(F^{k+1}(x_k) - F^{k+1}(x^*)) \\
&\quad \leq \frac{1}{2\gamma_{k+1}}\left(\|u_k - x^*\|^2 - \mathbb{E}_k\left[\|u_{k+1} - x^*\|^2\right]\right) + t_{k+1}^2\gamma_{k+1}\left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right).
\end{aligned} \tag{3.27}$$

Next, by adding $t_k^2(F^{k+1}(x_k) - F^{k+1}(x^*))$ on both sides of (3.27), gives

$$\begin{aligned}
&\mathbb{E}_k\left[t_{k+1}^2 \Delta_{k+1}\right] + \rho_{k+1}(F^{k+1}(x_k) - F^{k+1}(x^*)) + \frac{1}{2\gamma_{k+1}}\mathbb{E}_k\left[\|u_{k+1} - x^*\|^2\right] \\
&\quad \leq t_k^2(F^{k+1}(x_k) - F^{k+1}(x^*)) + \frac{1}{2\gamma_{k+1}}\|u_k - x^*\|^2 + t_{k+1}^2\gamma_{k+1}\left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right).
\end{aligned}$$

Using (3.5) together with the assumption that $(\lambda_k)_{k\geq 1}$ is nonincreasing leads to

$$\begin{aligned}
&\mathbb{E}_k\left[t_{k+1}^2 \Delta_{k+1}\right] + \rho_{k+1}(F^{k+1}(x_k) - F^{k+1}(x^*)) + \frac{1}{2\gamma_{k+1}}\mathbb{E}_k\left[\|u_{k+1} - x^*\|^2\right] \\
&\quad \leq t_k^2\Delta_k + \frac{1}{2\gamma_{k+1}}\|u_k - x^*\|^2 + t_k^2(\lambda_k - \lambda_{k+1})\frac{L_g^2}{2} + t_{k+1}^2\gamma_{k+1}\left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right).
\end{aligned}$$

Using that $t_k^2 \geq t_{k+1}^2 - t_{k+1}$, we get

$$\begin{aligned}
&\mathbb{E}_k\left[t_{k+1}^2 \Delta_{k+1}\right] + \rho_{k+1}(F^{k+1}(x_k) - F^{k+1}(x^*)) \\
&\quad \leq t_k^2\Delta_k + \frac{1}{2\gamma_{k+1}}(\|u_k - x^*\|^2 - \mathbb{E}_k\left[\|u_{k+1} - x^*\|^2\right]) \\
&\qquad + t_k^2\lambda_k\frac{L_g^2}{2} - t_{k+1}^2\lambda_{k+1}\frac{L_g^2}{2} + t_{k+1}\lambda_{k+1}\frac{L_g^2}{2} + t_{k+1}^2\gamma_{k+1}\left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right).
\end{aligned}$$

Multiplying both sides with γ_{k+1} and putting all terms on the correct sides yields

$$\begin{aligned}
&\mathbb{E}_k\left[\gamma_{k+1}t_{k+1}^2\left(\Delta_{k+1} + \lambda_{k+1}\frac{L_g^2}{2}\right) + \frac{1}{2}\|u_{k+1} - x^*\|^2\right] + \gamma_{k+1}\rho_{k+1}(F^{k+1}(x_k) - F^{k+1}(x^*)) \\
&\quad \leq \gamma_{k+1}t_k^2\left(\Delta_k + \lambda_k\frac{L_g^2}{2}\right) + \frac{1}{2}\|u_k - x^*\|^2 \\
&\qquad + \gamma_{k+1}t_{k+1}\lambda_{k+1}\frac{L_g^2}{2} + t_{k+1}^2\gamma_{k+1}^2\left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right).
\end{aligned} \tag{3.28}$$

At this point we would like to discard the term $\gamma_{k+1}\rho_{k+1}(F^{k+1}(x_k) - F^{k+1}(x^*))$ which we currently cannot as the positivity of $F^{k+1}(x_k) - F^{k+1}(x^*)$ is not ensured. So we add $\gamma_{k+1}\rho_{k+1}\lambda_{k+1}\frac{L_g^2}{2}$ on both sides of (3.28) and use the fact that (see (3.5))

$$\gamma_{k+1}\rho_{k+1}\left(F^{k+1}(x_k) - F^{k+1}(x^*) + \lambda_{k+1}\frac{L_g^2}{2}\right) \geq \gamma_{k+1}\rho_{k+1}(F(x_k) - F(x^*)) \geq 0$$

to deduce that

$$\begin{aligned}\mathbb{E}_k&\left[\gamma_{k+1}t_{k+1}^2\left(\Delta_{k+1}+\lambda_{k+1}\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_{k+1}-x^*\|^2\right]\\&\leq \gamma_{k+1}t_k^2\left(\Delta_k+\lambda_k\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_k-x^*\|^2\\&\quad+\gamma_{k+1}\lambda_{k+1}\frac{L_g^2}{2}(t_{k+1}+\rho_{k+1})+t_{k+1}^2\gamma_{k+1}^2\left(\sigma^2+\frac{1}{2}\|A\|^2L_g^2\right).\end{aligned}\tag{3.29}$$

Last but not least we use the that $\Delta_k+\lambda_k\frac{L_g^2}{2}\geq F(x_k)-F(x^*)\geq 0$ and $\gamma_{k+1}\leq\gamma_k$ to follow that

$$\gamma_{k+1}t_k^2\left(\Delta_k+\lambda_k\frac{L_g^2}{2}\right)\leq\gamma_k t_k^2\left(\Delta_k+\lambda_k\frac{L_g^2}{2}\right).\tag{3.30}$$

Combining (3.29) and (3.30) yields

$$\begin{aligned}\mathbb{E}_k&\left[\gamma_{k+1}t_{k+1}^2\left(\Delta_{k+1}+\lambda_{k+1}\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_{k+1}-x^*\|^2\right]\\&\leq \gamma_k t_k^2\left(\Delta_k+\lambda_k\frac{L_g^2}{2}\right)+\frac{1}{2}\|u_k-x^*\|^2\\&\quad+\gamma_{k+1}\lambda_{k+1}\frac{L_g^2}{2}(t_{k+1}+\rho_{k+1})+t_{k+1}^2\gamma_{k+1}^2\left(\sigma^2+\frac{1}{2}\|A\|^2L_g^2\right).\end{aligned}\tag{3.31}$$

Let $N\geq 2$. We take the expected value on both sides (3.31) and sum from $k=1$ to $N-1$. Getting rid of the non-negative terms $\|u_N-x^*\|^2$ gives

$$\begin{aligned}\gamma_N t_N^2\mathbb{E}\left[\Delta_N+\lambda_N\frac{L_g^2}{2}\right]&\leq\gamma_1\mathbb{E}\left[\Delta_1+\lambda_1\frac{L_g^2}{2}\right]+\frac{1}{2}\|u_1-x^*\|^2\\&\quad+\sum_{k=2}^{N}\gamma_k\lambda_k\frac{L_g^2}{2}(t_k+\rho_k)\sum_{k=2}^{N}t_k^2\gamma_k^2\left(\sigma^2+\frac{1}{2}\|A\|^2L_g^2\right).\end{aligned}$$

Since $t_1=1$, the above inequality holds also for $N=1$. Now, using Lemma 3.4.3 we get that for every $N\geq 1$

$$\gamma_N t_N^2\mathbb{E}\left[\Delta_N+\lambda_N\frac{L_g^2}{2}\right]\leq\ \frac{1}{2}\|x_0-x^*\|^2+\sum_{k=1}^{N}\gamma_k\lambda_k\frac{L_g^2}{2}(t_k+\rho_k)+\sum_{k=1}^{N}t_k^2\gamma_k^2\left(\sigma^2+\frac{\|A\|^2}{2}\right).$$

From (3.6) we follow that

$$\gamma_N t_N^2\left(F(x_N)-F(x^*)\right)\leq\gamma_N t_N^2\left(\Delta_N+\lambda_N\frac{L_g^2}{2}\right),$$

therefore, for every $N \geq 1$

$$\gamma_N t_N^2 \mathbb{E}[F(x_N) - F(x^*)] \leq \frac{1}{2}\|x_0 - x^*\|^2 + \sum_{k=1}^{N} \gamma_k \lambda_k \frac{L_g^2}{2}(t_k + \rho_k) + \sum_{k=1}^{N} t_k^2 \gamma_k^2 \left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right).$$

By using the fact that $\lambda_k = \gamma_k \|A\|^2$ for every $k \geq 1$ we deduce by dividing by $\gamma_N t_N^2$ that

$$\mathbb{E}[F(x_N) - F(x^*)] \leq \frac{1}{\gamma_N t_N^2}\frac{1}{2}\|x_0 - x^*\|^2 + \frac{1}{\gamma_N t_N^2}\frac{\|A\|^2 L_g^2}{2}\sum_{k=1}^{N}\gamma_k^2(t_k + \rho_k) + \frac{1}{\gamma_N t_N^2}\left(\sigma^2 + \frac{1}{2}\|A\|^2 L_g^2\right)\sum_{k=1}^{N} t_k^2 \gamma_k^2 \quad \forall N \geq 1.$$

□

Corollary 3.4.5. *Let*

$$t_1 = 1, \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \quad \forall k \geq 1,$$

and, for $b > 0$,

$$\lambda_k = \frac{b}{k^{\frac{3}{2}}}\|A\|^2, \text{ and } \gamma_k = \frac{b}{k^{\frac{3}{2}}} \quad \forall k \geq 1.$$

Then,

$$\mathbb{E}[F(x_N) - F(x^*)] \leq 2\frac{\|x_0 - x^*\|^2}{b\sqrt{N}} + b\|A\|^2 L_g^2 \frac{\pi^2}{3}\frac{1}{\sqrt{N}} + 2b\left(2\sigma^2 + \|A\|^2 L_g^2\right)\frac{1 + \log(N)}{\sqrt{N}} \quad \forall N \geq 1.$$

Furthermore, we have that $F(x_N)$ converges almost surely to $F(x^*)$ as $N \to +\infty$.

Proof. First we notice that the choice of $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ fulfills that

$$\rho_{k+1} = t_k^2 - t_{k+1}^2 + t_{k+1} = 0 \quad \forall k \geq 1.$$

Now we derive the stated convergence result by first showing via induction that

$$\frac{1}{k} \leq \frac{1}{t_k} \leq \frac{2}{k} \quad \forall k \geq 1.$$

Assuming that this holds for $k \geq 1$, we have that

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \leq \frac{1 + \sqrt{1 + 4k^2}}{2} \leq \frac{1 + \sqrt{1 + 4k + 4k^2}}{2} = k + 1$$

and

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \geq \frac{1 + \sqrt{1 + 4(\frac{k}{2})^2}}{2} \geq \frac{1 + \sqrt{k^2}}{2} \geq \frac{k+1}{2}.$$

Furthermore, for every $N \geq 1$ we have that

$$\begin{aligned} \frac{1}{\gamma_N t_N^2} \frac{\|A\|^2 L_g^2}{2} \sum_{k=1}^{N} \gamma_k^2 (t_k + \rho_k) &\leq \frac{4}{b\sqrt{N}} \frac{\|A\|^2 L_g^2}{2} \sum_{k=1}^{N} \frac{b^2}{k^3} k = \frac{2b\|A\|^2 L_g^2}{\sqrt{N}} \sum_{k=1}^{N} k^{-2} \\ &\leq \frac{2b\|A\|^2 L_g^2}{\sqrt{N}} \sum_{k=1}^{\infty} k^{-2} = b\|A\|^2 L_g^2 \frac{\pi^2}{3} \frac{1}{\sqrt{N}}. \end{aligned} \tag{3.32}$$

The statement of the convergence rate in expectation follows now by plugging in our parameter choices into the statement of Theorem 3.4.4, using the estimate (3.32) and checking that

$$\sum_{k=1}^{N} t_k^2 \gamma_k^2 \leq b^2 \sum_{k=1}^{N} \frac{1}{k} \leq b^2 (1 + \log(N)) \quad \forall N \geq 1.$$

The almost sure convergence of $(F(x_N))_{N \geq 1}$ can be deduced by looking at (3.31) and dividing by $\gamma_{k+1} t_{k+1}^2$ and using that $\gamma_{k+1} t_{k+1}^2 \geq \gamma_k t_k^2$ as well as $\rho_k = 0$, which gives for every $k \geq 0$

$$\begin{aligned} \mathbb{E}_k &\left[F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \lambda_{k+1} \frac{L_g^2}{2} + \frac{1}{2\gamma_{k+1} t_{k+1}^2} \|u_{k+1} - x^*\|^2 \right] \\ &\leq F^k(x_k) - F^k(x^*) + \lambda_k \frac{L_g^2}{2} + \frac{1}{2\gamma_k t_k^2} \|u_k - x^*\|^2 + \frac{\lambda_{k+1}}{t_{k+1}} \frac{L_g^2}{2} + \gamma_{k+1} \left(\sigma^2 + \frac{1}{2} \|A\|^2 L_g^2 \right). \end{aligned}$$

Plugging in our choice of parameters gives for every $k \geq 0$

$$\begin{aligned} \mathbb{E}_k &\left[F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \lambda_{k+1} \frac{L_g^2}{2} + \frac{1}{2\gamma_{k+1} t_{k+1}^2} \|u_{k+1} - x^*\|^2 \right] \\ &\leq F^k(x_k) - F^k(x^*) + \lambda_k \frac{L_g^2}{2} + \frac{1}{2\gamma_k t_k^2} \|u_k - x^*\|^2 + \frac{C}{k^{\frac{3}{2}}}, \end{aligned}$$

where $C > 0$.

Thus, by the famous Robbins-Siegmund Theorem (see [90, Theorem 1]) we get that $(F^{k+1}(x_{k+1}) - F^{k+1}(x^*) + \lambda_{k+1} \frac{L_g^2}{2})_{k \geq 0}$ converges almost surely. In particular, from the convergence to 0 in expectation we know that the almost sure limit must also be the constant zero. □

Finite sum. The formulation of the previous section can be used to deal e.g. with problems of the form

$$\min_{x \in \mathbb{R}^d} f(x) + \sum_{i=1}^{m} g_i(A_i x) \tag{3.33}$$

for $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ a proper, convex and lower semicontinuous function, $g_i : \mathbb{R}^n_i \to \mathbb{R}$ convex and L_{g_i}-Lipschitz continuous functions and $A_i : \mathbb{R}^d \to \mathbb{R}^n_i$ linear continuous operators for $i = 1, \ldots, m$.

Clearly one could consider

$$\boldsymbol{A} := \begin{cases} \mathbb{R}^d \to \times_{i=1}^m \mathbb{R}^n_i \\ x \mapsto \times_{i=1}^m A_i x \end{cases}$$

with $\|\boldsymbol{A}\|^2 = \sum_{i=1}^m \|A_i\|^2$ and

$$\boldsymbol{g} := \begin{cases} \times_{i=1}^m \mathbb{R}^n_i \to \mathbb{R} \cup \{+\infty\} \\ \times_{i=1}^m y_i \mapsto \sum_{i=1}^m g_i(y_i) \end{cases}$$

in order to reformulate the problem as

$$\min_{x \in \mathbb{R}^d} f(x) + \boldsymbol{g}(\boldsymbol{A}x)$$

and use Algorithm 3.3.1 together with the parameter choices described in Corollary 3.3.6 on this. This results in the following algorithm.

Algorithm 3.4.6. Let $y_0 = x_0 \in \mathbb{R}^d, \lambda_1 = b\|\boldsymbol{A}\|$, for $b > 0$, and $t_1 = 1$. Consider the following iterative scheme

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} \gamma_k = \frac{\sum_{i=1}^m \|A_i\|^2}{\lambda_k} \\ x_k = \operatorname{prox}_{\gamma_k f}\left(y_{k-1} - \gamma_k \sum_{i=1}^m A_i^* \operatorname{prox}_{\frac{1}{\lambda_k} g_i^*}\left(\frac{A_i y_{k-1}}{\lambda_k}\right)\right) \\ t_{k+1} = \sqrt{t_k^2 + 2t_k} \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \\ \lambda_{k+1} = \lambda_k \frac{t_k^2}{t_{k+1}^2 - t_{k+1}}. \end{array} \right.$$

However, problem (3.33) also lends itself to be tackled via the stochastic version of our method, Algorithm 3.4.1, by randomly choosing a subset of the summands. Together with the parameter choices described in Corollary 3.4.5 which results in the following scheme.

Algorithm 3.4.7. Let $y_0 = x_0 \in \mathbb{R}^d, b > 0$, and $t_1 = 1$. Consider the following iterative scheme

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} \lambda_k = b \sum_{i=1}^m \|A_i\|^2 k^{-\frac{3}{2}} \\ \gamma_k = b k^{-\frac{3}{2}} \\ x_k = \operatorname{prox}_{\gamma_k f}\left(y_{k-1} - \gamma_k \frac{\epsilon_{i,k}}{p_i} \sum_{i=1}^m A_i^* \operatorname{prox}_{\frac{1}{\lambda_k} g_i^*}\left(\frac{A_i y_{k-1}}{\lambda_k}\right)\right) \\ t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ y_k = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}), \end{array} \right.$$

with $\epsilon_k := (\epsilon_{1,k}, \epsilon_{2,k}, \ldots, \epsilon_{m,k})$ a sequence of i.i.d., $\{0,1\}^m$ random variables and $p_i = \mathbb{P}[\epsilon_{i,1} = 1]$.

Since the above two methods were not explicitly developed for this separable case and can therefore not make use of more refined estimation of the constant $\|\boldsymbol{A}\|$, as it is done in e.g. [27]. However, in the stochastic case, this fact is remedied due to the scaling of the stepsize with respect to the i-th component by p_i^{-1}.

Remark 3.4.2. In theory Algorithm 3.4.1 could be used to treat more general stochastic problems than finite sums like (3.33), but in the former case it is not clear anymore how a gradient estimator can be found, so we do not discuss it here.

3.5 Numerical experiments

We will focus our numerical experiments on image processing problems. The examples are implemented in python using the operator discretization library (ODL) [1]. We define the discrete gradient operators D_1 and D_2 representing the discretized derivative in the first and second coordinate respectively, which we will need for the numerical examples. Both map from $\mathbb{R}^{r\times s}$ to $\mathbb{R}^{r\times s}$ and are defined by

$$(D_1 u)_{i,j} := \begin{cases} u_{i+1,j} - u_{i,j} & 1 \le i < m, \\ 0 & \text{else,} \end{cases}$$

and

$$(D_2 u)_{i,j} := \begin{cases} u_{i,j+1} - u_{i,j} & 1 \le j < m, \\ 0 & \text{else.} \end{cases}$$

The operator norm of D_1 and D_2, respectively, is 2 (where we equipped $\mathbb{R}^{r\times s}$ with the Frobenius norm). This yields an operator norm of $\sqrt{8}$ for the total gradient $D := D_1 \times D_2$ as a map from $\mathbb{R}^{r\times s}$ to $\mathbb{R}^{r\times s} \times \mathbb{R}^{r\times s}$, see also [25].

We will compare our methods, i.e. the Variable Accelerated SmooThing (VAST) and its stochastic counterpart (sVAST) to the Primal Dual Hybrid Gradient (PDHG) of [28] as well as its stochastic version (sPDHG) from [27]. Furthermore, we will illustrate another competitor, the method by Pesquet and Repetti, see [83], which is another stochastic version of PDHG (see also [105]).

In all examples we choose the parameters in accordance with [27]:

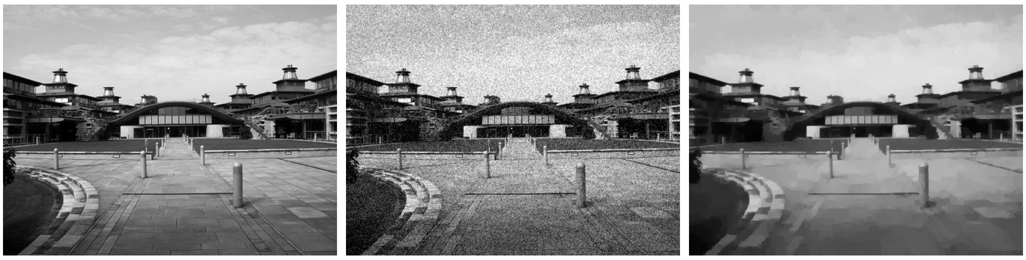
- for PDHG and Pesquet&Repetti: $\tau = \sigma_i = \frac{\gamma}{\|A\|}$
- for sPDHG: $\sigma_i = \frac{\gamma}{\|A\|}$ and $\tau = \frac{\gamma}{n \max_i \|A_i\|}$,

where $\gamma = 0.99$.

3.5.1 Total variation denoising

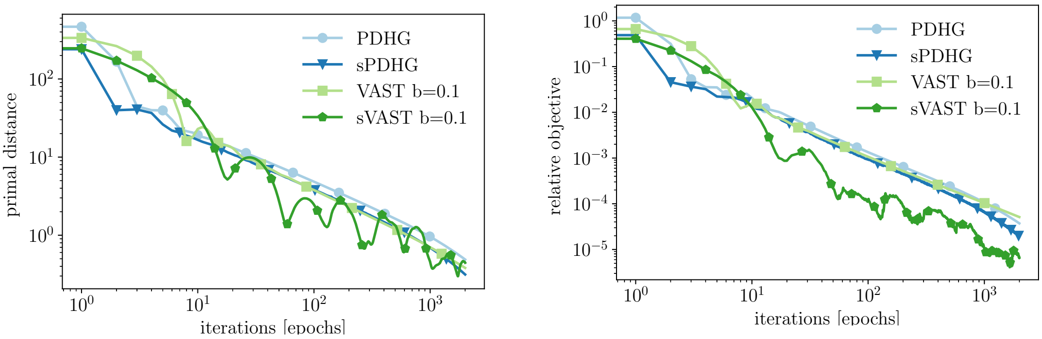
The task at hand is to reconstruct an image from its noisy observation. We do this by solving

$$\min_{x\in\mathbb{R}^{r\times s}} \alpha\|x-b\|_2 + \|D_1 x\|_1 + \|D_2 x\|_1, \tag{3.34}$$



(a) Groundtruth (b) Data (c) Approximate solution

Figure 3.1: TV denoising. Images used. The approximate solution is computed by running PDHG for 7000 iterations.



(a) Distance to the solution. (b) Relative objective $\frac{F(x_k)-F(x^*)}{F(x_0)-F(x^*)}$.

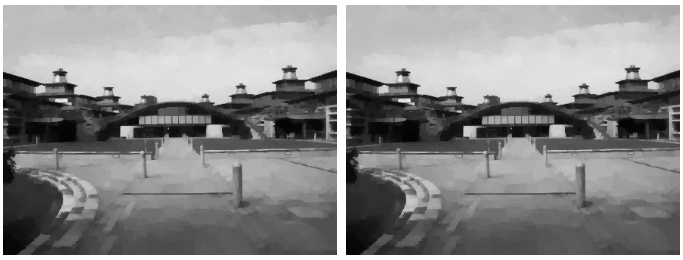
Figure 3.2: A comparison of different methods on the problem of TV denoising.

with $\alpha > 0$ as regularization parameter, in the following setting: $f = \alpha\|\cdot - b\|_2, g_1 = g_2 = \|\cdot\|_1, A_1 = D_1, A_2 = D_2$.

Figure 3.1 illustrates the images (of dimension $r = 442$ and $s = 331$) used in for this example. These include the groundtruth, i.e. the uncorrupted image, as well as the data for the optimization problem b, which visualizes the level of noise. In Figure 3.2 we can see that for the deterministic setting our method is as good as PDHG. For the objective function values, Subfigure 3.2b, this is not too surprising as both algorithms share the same convergence rate. For the distance to a solution however we completely lack a convergence result. Nevertheless in Subfigure 3.2a we can see that our method performs also well with respect to this measure.

In the stochastic setting we can see in Figure 3.2 that, while sPDHG provides some benefit over its deterministic counterpart, the stochastic version of our method, although significantly increasing the variance, provides great benefit, at least for the objective function values.

Furthermore, Figure 3.3, shows the reconstructions of sPDHG and our method which are, despite the different objective function values, quite comparable.



(a) sVAST (b) sPDHG

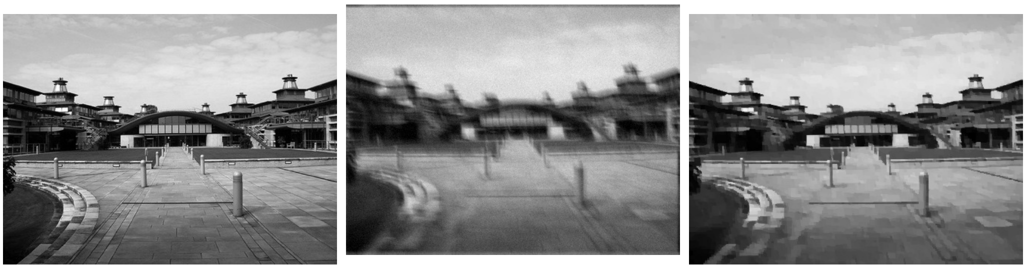
Figure 3.3: TV Denoising. A comparison of the reconstruction for the stochastic variable smoothing method and the stochastic PDHG.

3.5.2 Total variation deblurring

For this example we want to reconstruct an image from a blurred and noisy image. We assume to know the blurring operator $C : \mathbb{R}^{r\times s} \to \mathbb{R}^{r\times s}$. This is done by solving

$$\min_{x\in\mathbb{R}^{r\times s}} \alpha\|Cx - b\|_2+\|D_1x\|_1+\|D_2x\|_1, \tag{3.35}$$

for $\alpha > 0$ as regularization parameter, in the following setting: $f = 0, g_1 = \alpha\|\cdot-b\|_2, g_2 = g_3 = \|\cdot\|_1, A_1 = C, A_2 = D_1, A_2 = D_2$.

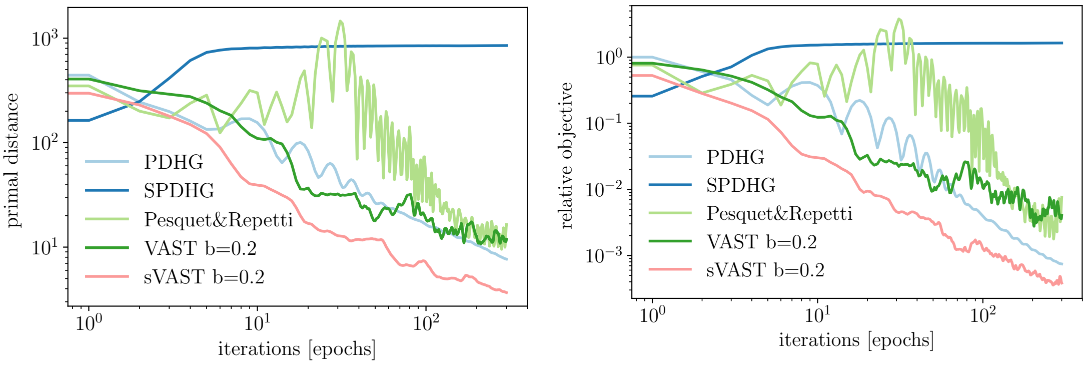


(a) Groundtruth (b) Data (c) Approximate solution

Figure 3.4: TV Deblurring.The approximate solution is computed by running PDHG for 3000 iterations.

Figure 3.4 shows the images used to set up the optimization problem (3.35), in particular Subfigure 3.4b which corresponds to b in said problem.

In Figure 3.5 we see that while PDGH performs better in the deterministic setting, in particular in the later iteration, the stochastic variable smoothing method provides a significant improvement where sPDHG method seems not to converge. It is interesting to note that in this setting even the deterministic version of our algorithm exhibits a slightly chaotic behaviour. Although neither of the two methods is monotone in the primal objective function PDHG seems here much more stable.



(a) Distance to the solution.

(b) Relative objective $\frac{F(x_k)-F(x^*)}{F(x_0)-F(x^*)}$.

Figure 3.5: A comparison of different methods on the problem of TV deblurring.

3.5.3 Matrix factorization

In this section we want to solve a *nonconvex* and nonsmooth optimization problem of completely positive matrix factorization, see [30, 44]. For an observed matrix $M \in \mathbb{R}^{s\times s}$ we want to find a completely positive low rank factorization, meaning we are looking for $x \in \mathbb{R}^{r\times s}_{\geq 0}$ with $r \ll s$ such that $x^T x = M$. This can be formulated as the following (robust) optimization problem

$$\min_{x\in\mathbb{R}^{r\times s}_{\geq 0}} \|x^T x - M\|_1, \tag{3.36}$$

where x^T denotes the transpose of the matrix x. The more natural approach might be to use a smooth formulation where $\|\cdot\|_2^2$ is used instead of the 1-Norm we are suggesting. However, the former choice of distance measure, albeit smooth, comes with its own set of problems (mainly a non-Lipschitz gradient).

The so called *prox-linear method* presented in [37], solves the above problem (3.36), by linearizing the smooth ($\mathbb{R}^{s\times s}$-valued) function $x \mapsto x^T x$ inside the nonsmooth distance function. In particular for the problem

$$\min_x g(c(x))$$

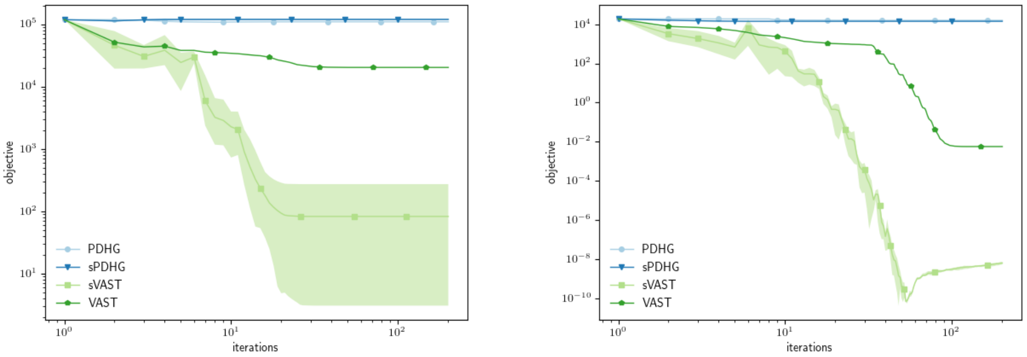
for a smooth vector valued function c and a convex and Lipschitz function g, [37] proposes to iteratively solve the subproblem

$$x_{k+1} = \arg\min_x \left\{ g\Big(c(x_k) + (Dc)(x_k)(x - x_k)\Big) + \frac{1}{2t}\|x - x_k\|_2^2 \right\} \tag{3.37}$$

for a stepsize $t \leq (L_g L_{Dc})^{-1}$. For our particular problem described in (3.36) the subproblem looks as follows

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^{r\times s}_{\geq 0}} \left\{ \|x_k^T x - M\|_1 + \frac{1}{2}\|x - x_k\|_2^2 \right\}, \tag{3.38}$$

Objective function in completely positive matrix factorization



(a) Random starting point.

(b) Starting point close to the solution.

Figure 3.6: Comparison of the for different starting points. We run 40 epochs with 5 iterations each. For each epoch we choose the last iterate of the previous epoch as the linearization. For the stochastic methods we fix the number of rows (batch size) which are randomly chosen in each update a priori and count d divided by this number as one iteration. For the randomly chosen initial point we use a batch size of 3 (to allow for more exploration) and for the one close to the solution we use 5 in order to give a more accuracy. The parameter b in the variable smoothing method was chosen with minimal tuning to be 0.1 for both the deterministic and the stochastic version.

and therefore fits our general setup described in (3.1) with the identification $f = \|\cdot - x_k\|_2^2 + \delta_{\mathbb{R}^{r\times s}_{\geq 0}}(x)$, $g = \|\cdot\|_1$ and $A = x_k^T$. Moreover, due to its separable structure, the subproblem (4.3) fits the special case described in (3.33) and can therefore be tackled by the stochastic version of our algorithm presented in Algorithm 3.4.7. In particular reformulating (3.37) for the stochastic finite sum setting we interpret the subproblem as

$$x_{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^{r\times s}_{\geq 0}} \left\{ \sum_{i=1}^{s} \left\| x_k^T[i,:]x - M[i,:] \right\|_1 + \frac{1}{2}\|x - x_k\|_2^2 \right\}, \tag{3.39}$$

where $M[i,:]$ denotes the i-th row of the matrix M.

In comparison to Section 3.5.1 and Section 3.5.2 a new aspect becomes important when evaluating methods for solving (3.37). Now, it is not only relevant how well subproblem (4.3) is solved but also the trajectory taken in doing so as different paths might lead to different local minima. This can be seen in Figure 3.6 where PDHG gets stuck early on in bad local minima. The variable smoothing method (especially the stochastic version) is able to move further from the starting point and find better local minima. Note that in general the methods have a difficulty finding the global minimum $x_{true} \in \mathbb{R}^{3\times 60}$ (with optimal objective function value zero, as constructed $M := x_{true}^T x_{true} \in \mathbb{R}^{60\times 60}$ in all examples).

4 Variable smoothing for weakly convex composite problems

We study minimization of a structured objective function, being the sum of a smooth function and a composition of a weakly convex function with a linear operator. Applications include image reconstruction problems with regularizers that introduce less bias than the standard convex regularizers. We develop a variable smoothing algorithm, based on the Moreau envelope with a decreasing sequence of smoothing parameters, and prove a complexity of $\mathcal{O}(\epsilon^{-3})$ to achieve an ϵ-approximate solution. This bound interpolates between the $\mathcal{O}(\epsilon^{-2})$ bound for the smooth case and the $\mathcal{O}(\epsilon^{-4})$ bound for the subgradient method. Our complexity bound is in line with other works that deal with structured nonsmoothness of weakly convex functions.

4.1 Problem setting and motivation

We study minimization of the sum of a smooth function h and a possibly nonsmooth, weakly convex function g composed with a linear operator defined by the matrix $A \in \mathbb{R}^{n\times d}$, that is,

$$\min_{x\in\mathbb{R}^d} \Big\{F(x) := h(x) + g(Ax)\Big\}. \tag{4.1}$$

Our approach makes use of the *Moreau envelope* (see Definition 2.3.6) g_λ, for a positive scalar λ, together with gradient descent. Steps of the algorithm have the form

$$x \leftarrow x - \gamma\nabla(h + g_\lambda \circ A)(x),$$

for some step length γ. For accelerated versions of this approach for convex problems see Chapter 3, or [23, 101].

4.1.1 Composite problems

We discuss several nonconvex instances of problems of the form (4.1) as the convex case has been discussed in Chapter 3.

Weakly Convex Regularizers. Functions that are "sharp" around zero have a long history as sparsity-inducing regularizers. Foremost among such functions is the ℓ_1 norm $\|\cdot\|_1$, which is used for example in sparse least-squares regression (also known as LASSO):

$$\min_x \frac{1}{2}\|Ax - b\|^2 + \|x\|_1$$

or the anisotropic Total Variation denoising or deblurring problems (3.34) and (3.35). However, the use of the ℓ_1 regularizer tends to depress the magnitude of nonzero elements of the solution, resulting in *bias.* This phenomenon is a consequence of the fact that the proximal operator of the 1-norm, often called the *soft thresholding operator*, does not approach the identity for larger values of its argument. For this reason, nonconvex alternatives to $\|\cdot\|_1$ are often used to reduce bias. These include ℓ_p-norms (with $0 < p < 1$) which are not weakly convex, and the several weakly convex regularizers, which we now describe. The *minimax concave penalty (MCP)*, introduced in [108] and used in [57, 96], is a family of functions $r_{\nu,\theta} : \mathbb{R} \to \mathbb{R}_+$ involving two positive parameters ν and θ, and defined by

$$r_{\nu,\theta}(x) := \begin{cases} \nu|x| - \frac{x^2}{2\theta}, & |x| \leq \theta\nu, \\ \frac{\theta\nu^2}{2}, & \text{otherwise.} \end{cases}$$

(Note that this function satisfies the definition of ρ-weak convexity with $\rho = \theta^{-1}$.) The proximal operator of this function (called *firm threshold* in [7]) can be written in the following closed form when $\theta > \beta$:

$$\operatorname{prox}_{\beta r_{\nu,\theta}}(x) = \begin{cases} 0, & |x| < \beta\nu, \\ \frac{x - \nu\beta \operatorname{sgn}(x)}{1 - (\beta/\theta)}, & \beta\nu \leq |x| \leq \theta\nu, \\ x, & |x| > \theta\nu. \end{cases}$$

The *fractional penalty function* (cf. [57, 82]) $\varphi_a : \mathbb{R} \to \mathbb{R}_+$ (for parameter $a > 0$) is

$$\varphi_a(x) := \frac{|x|}{1 + \frac{a}{2}|x|}.$$

The *smoothly clipped absolute deviation (SCAD)* [38] (cf. [57]) is defined for parameters $\nu > 0$ and $\theta > 2$ as follows:

$$r_{\nu,\theta}(x) = \begin{cases} \nu|x|, & |x| \leq \nu, \\ \frac{-x^2 + 2\theta\nu|x| - \nu^2}{2(\theta - 1)}, & \nu < |t| \leq \theta\nu, \\ \frac{(\theta+1)\nu^2}{2}, & |t| > \theta\nu. \end{cases}$$

(This function is $(\theta - 1)^{-1}$-weakly convex.)

Since these functions approach (or attain) a finite value as their argument grows in magnitude, they do not introduce as much bias in the solution as does the ℓ_1 norm, and their proximal operators approach the identity for large arguments.

These regularizers have, however, mostly been used in the simple additive setting

$$\min_{x \in \mathbb{R}^d} h(x) + g(x)$$

for a smooth data fidelity term h and nonsmooth regularizer g, for example in least squares or logistic regression [96] and compressed sensing (cf. [7]).

Weakly convex composite losses. The use of weakly convex functions composed with linear operators has been explored in the robust statistics literature. An early instance is the *Tukey biweight* function [8], in which $g(Ax)$ has the form

$$g(Ax) = \sum_{i=1}^{n} \varphi(A_{i}.x - b_i), \quad \text{where } \varphi(\theta) = \frac{\theta^2}{1+\theta^2}. \tag{4.2}$$

This function behaves like the usual least-squares loss when $\theta^2 \ll 1$ but asymptotes at 1. It is ρ-weakly convex with $\rho = 6$.

A different (but similar) definition of the Tukey biweight function appears in [63, Section 2.1]. This same reference also mentions another nonconvex loss function, the *Cauchy loss*, which has the form (4.2) except that φ is defined by

$$\varphi(\theta) = \frac{\xi^2}{2} \log\left(1 + \frac{\theta^2}{\xi^2}\right),$$

for some parameter ξ. This function is ρ-weakly convex with $\rho = 6$.

4.1.2 Complexity bounds for weakly convex problems

To put our results in perspective, we provide a review of the literature on complexity bounds for optimization problems related to our formulation (4.1), including weakly convex functions. In all cases, these are bounds on the number of iterations required to find an approximately stationary point, where our measure of stationarity is based the norm of the gradient of the Moreau envelope (a smooth proxy).

The best known complexity for black box subgradient optimization for weakly convex functions is $\mathcal{O}(\epsilon^{-4})$. This result is proved for *stochastic* subgradients in [35], but as in the convex case, there is no known improvement in the deterministic setting. As in convex optimization, subgradient methods are quite general and implementable for weakly convex functions. However, when more structure is present in the function, algorithms that achieve better complexity can be devised. In particular, when the proximal operator of the nonsmooth weakly convex function can be calculated analytically, complexity bounds of $\mathcal{O}(\epsilon^{-2})$ can be proven (see Section 4.3), the same bounds as for steepest descent methods in the smooth nonconvex case. This means that the entire difficulty introduced by the nonsmoothness can be mitigated as long as it is treated by a proximal operator.

For convex optimization problems, bounds of $\mathcal{O}(\epsilon^{-1})$ can be obtained for gradient methods on smooth functions and $\mathcal{O}(\epsilon^{-1/2})$ for accelerated gradient methods. These same bounds can also be obtained for nonsmooth problems provided that the nonsmooth function is handled by a proximal operator. When the explicit proximal operator is not available and subgradient methods have to be used, the complexity reverts to $\mathcal{O}(\epsilon^{-2})$.

It is possible to keep the $\mathcal{O}(\epsilon^{-2})$ rate when just a local model of the weakly convex part is evaluated by a convex operator. The paper [37] studies optimization problems of the type

$$\min_x \; h(x) + g(c(x))$$

where h is proper, convex and lower semicontinuous, g is convex and Lipschitz continuous, and c is smooth. (Under these assumptions, the composition $g \circ c$ is weakly convex.) The $\mathcal{O}(\epsilon^{-2})$ bound is proved for an algorithm in which the (convex) subproblem

$$\min_{y} \; h(y) + g(c(x) + \nabla c(x)(y - x)) + \frac{1}{2t}\|y - x\|^2 \tag{4.3}$$

is solved explicitly. In the more realistic case in which (4.3) must be solved by an iterative procedure, a bound of $\tilde{\mathcal{O}}(\epsilon^{-3})$ is obtained in [37].

Functions of the form $g(c(x))$ have also been studied in [56] for the case of a smooth nonlinear vector function c and a prox-regular g. This is more general than the formulations consider in this paper, both in the fact that all weakly convex functions are prox-regular, and in the nonlinearity of the inner map. The subproblems in [56] have a form similar to (4.3), and while convergence results are proved in the latter paper, it does not contain rate-of-convergence results or complexity results.

A different weakly convex structure is explored in Chapter 6, in which the weak convexity stems from a smooth saddle point problem. We consider there

$$\min_{x} \max_{y} \; \Phi(x, y),$$

where $\Phi(x, \cdot)$ is concave, $\Phi(\cdot, y)$ is nonconvex, and $\Phi(\cdot, \cdot)$ is smooth. In this setting a bound of $\tilde{\mathcal{O}}(\epsilon^{-3})$ for a method that uses only gradient evaluations can be achieved [61,99].

In light of the considerations above, the complexity bound of $\mathcal{O}(\epsilon^{-3})$ for our algorithm seems almost inevitable. It interpolates between the setting without structural assumptions about the nonsmoothness (black box subgradient) and the perfect structural knowledge of the nonsmoothness (explicit knowledge of the proximal operator).

In Section 4.3, we treat the simpler setting in which the linear operator from (4.1) is the identity, so that $F(x) = h(x) + g(x)$. Similar problems have been analyzed before, for example, in [7, 96]. However, it is assumed in [7] that convexity in the data fidelity term h compensates for nonconvexity in the regularizer g such that the overall objective function F remains convex. (We make no such assumption here.) The paper [96] does not make such restrictive assumptions and proves convergence but not complexity bounds.

4.1.3 Stationarity

Recall that we say that a point x^* is a stationary point for a function if the Fréchet subdifferential of the function contains 0 at x^*. The concept of *nearly stationary* is not quite so straightforward. We motivate our approach by looking first at the simple additive composite problem, also discussed in Section 4.3, which corresponds to setting $A = I$ in (4.1), that is,

$$\min_{x} \; h(x) + g(x). \tag{4.4}$$

Stationarity for (4.4) means that $0 \in \partial(h+g)(x^*)$, that is, $-\nabla h(x^*) \in \partial g(x^*)$. A natural definition for ϵ-approximate stationarity for a point x would thus be

$$\operatorname{dist}(-\nabla h(x), \partial g(x)) \leq \epsilon.$$

However, since we are running gradient descent on the *smoothed* problem, our algorithm will naturally compute and detect points with that satisfy a threshold condition of the form

$$\|\nabla h(x) + \nabla g_\lambda(x)\| \le \epsilon. \tag{4.5}$$

Recall Lemma 2.3.8 which says that $\nabla g_\lambda(x) \in \partial g(\operatorname{prox}_{\lambda g}(x))$ for all x. It tells us that when (4.5) holds, then

$$\operatorname{dist}(-\nabla h(x), \partial g(\operatorname{prox}_{\lambda g}(x))) \le \epsilon,$$

which means that the two arguments of ∇h and ∂g do not quite match. In general, $\operatorname{prox}_{\lambda g}(x)$ might even be arbitrarily far away from x. We can remedy this issue by requiring g to be Lipschitz continuous, see Lemma 2.3.9.

When $x \in \mathbb{R}^d$ satisfies (4.5), ∇h is $L_{\nabla h}$-Lipschitz and g is L_g-Lipschitz, we have

$$\begin{aligned}
&\operatorname{dist}(-\nabla h(\operatorname{prox}_{\lambda g}(x)), \partial g(\operatorname{prox}_{\lambda g}(x))) \\
&\quad \le \|\nabla h(\operatorname{prox}_{\lambda g}(x)) - \nabla h(x)\| + \operatorname{dist}(-\nabla h(x), \partial g(\operatorname{prox}_{\lambda g}(x))) \\
&\quad \le L_{\nabla h}\|x - \operatorname{prox}_{\lambda g}(x)\| + \epsilon && \text{(from (4.5) and (2.3))} \\
&\quad \le L_{\nabla h} L_g \lambda + \epsilon && \text{(from (2.5)).}
\end{aligned}$$

Thus, if λ is sufficiently small and x satisfies (4.5), then $\operatorname{prox}_{\lambda g}(x)$ is near-stationary for (4.4).

4.1.4 Stationarity for the composite problem

It follows immediately from (2.3) in Lemma 2.3.8 that for $\lambda \in (0, \rho^{-1})$, we have for all $x \in \mathbb{R}^d$

$$\nabla(g_\lambda \circ A)(x) = A^* \nabla g_\lambda(Ax) \in A^* \partial g(\operatorname{prox}_{\lambda g}(Ax)). \tag{4.6}$$

Extending the results of the previous section to the case of a general linear operator A in (4.1) requires some work. Stationarity for (4.1) requires that $0 \in \nabla h(x) + A^* \partial g(Ax)$, so ϵ-near stationarity requires

$$\operatorname{dist}(-\nabla h(x), A^* \partial g(Ax)) \le \epsilon. \tag{4.7}$$

Our method can compute a point x such that

$$\|\nabla h(x) + \nabla(g_\lambda \circ A)(x)\| \le \epsilon,$$

which by (4.6) implies that

$$\operatorname{dist}(-\nabla h(x), A^* \partial g(z)) \le \epsilon, \quad \text{for } z = \operatorname{prox}_{\lambda g}(Ax), \tag{4.8}$$

where, provided that g is L_g-Lipschitz continuous, we have

$$\|Ax - z\| \le L_g \lambda. \tag{4.9}$$

The bound in (4.8) measures the criticality, while the bound in (4.9) concerns feasibility. The bounds (4.8), (4.9) are not a perfect match with (4.7), since the subdifferentials of h and $g \circ A$ are evaluated at different points.

Surjectivity of A. When A is surjective, we can perturb the x that satisfies (4.8), (4.9) to a nearby point x^* that satisfies a bound of the form (4.7). Since $z = \operatorname{prox}_{\lambda g}(Ax)$ is in the range of A, we can define

$$x^* := \operatorname*{arg\,min}_{x' \in \mathbb{R}^d} \left\{ \|x - x'\|^2 : \; Ax' = z \right\}, \tag{4.10}$$

which is given explicitly by

$$x^* = x - A^*(AA^*)^{-1}(Ax - z) = x - A^\dagger(Ax - z)$$

where $A^\dagger := A^*(AA^*)^{-1}$ is the pseudoinverse of A. The operator norm of the pseudoinverse is bounded by the inverse of the smallest singular value $\sigma_{\min}(A)$ of A, so when g is L_g-Lipschitz continuous, we have from (4.9) that

$$\|x - x^*\| \le \sigma_{\min}(A)^{-1}\|Ax - z\| \le \sigma_{\min}(A)^{-1} L_g \lambda. \tag{4.11}$$

The point x^* is approximately stationary in the sense of (4.7), for λ sufficiently small, because

$$\begin{aligned}
&\operatorname{dist}(-\nabla h(x^*), A^*\partial g(Ax^*)) \\
&\quad \le \|\nabla h(x^*) - \nabla h(x)\| + \operatorname{dist}(-\nabla h(x), A^*\partial g(z)) && (\text{since } Ax^* = z = \operatorname{prox}_{\lambda g}(Ax)) \\
&\quad \le L_{\nabla h}\|x - x^*\| + \epsilon && (\text{from (4.8)}) \\
&\quad \le L_{\nabla h}\sigma_{\min}(A)^{-1} L_g \lambda + \epsilon && (\text{from (4.11)}).
\end{aligned} \tag{4.12}$$

By choosing λ small, x^* will be an approximate solution in the stronger sense (4.7) and not just the weaker notion of (4.8), (4.9), which we have to settle for if A is not surjective.

4.2 Main results

We describe our variable smoothing approaches for the problem (4.1), where we assume that h is $L_{\nabla h}$-smooth, g is possibly nonsmooth, ρ-weakly convex, and L_g-Lipschitz continuous, and A is a nonzero linear continuous operator. For convenience, we define the smoothed approximation $F_k : \mathbb{R}^d \to \mathbb{R}$ based on the Moreau envelope with parameter λ_k as follows:

$$F_k(x) := h(x) + g_{\lambda_k}(Ax).$$

We note from Lemma 2.3.7 and the chain rule that

$$\nabla F_k(x) = \nabla h(x) + \frac{1}{\lambda_k} A^*(Ax - \operatorname{prox}_{\lambda_k g}(Ax)). \tag{4.13}$$

The quantity L_k defined by

$$L_k := L_{\nabla h} + \|A\|^2 \max\left\{ \lambda_k^{-1}, \frac{\rho}{1 - \rho\lambda_k} \right\} \tag{4.14}$$

is a Lipschitz constant of the gradient of ∇F_k, see Lemma 2.3.7. When $\rho\lambda_k \le 1/2$, the maximum in (4.14) is achieved by λ_k^{-1}, so in this case we can define

$$L_k := L_{\nabla h} + \|A\|^2/\lambda_k. \tag{4.15}$$

4.2.1 An elementary approach

Our first algorithm takes gradient descent steps on the smoothed problem, that is,

$$x_{k+1} = x_k - \gamma_k \nabla F_k(x_k), \tag{4.16}$$

for certain values of the parameter λ_k and stepsize γ_k. From (4.13), the formula (4.16) is equivalent to

$$x_{k+1} = x_k - \frac{\gamma_k}{\lambda_k} A^*(Ax_k - \operatorname{prox}_{\lambda_k g}(Ax_k)) - \gamma_k \nabla h(x_k).$$

Our basic algorithm is described next.

Algorithm 4.2.1 (Variable smoothing)**.** For an initial value $x_1 \in \mathbb{R}^d$ we iterate

$$(\forall k \geq 1) \quad \left\lfloor \begin{array}{l} \lambda_k = (2\rho)^{-1} k^{-1/3}, \text{define } L_k \text{ as in (4.15), set } \gamma_k = 1/L_k \\ x_{k+1} = x_k - \gamma_k \nabla F_k(x_k) \end{array} \right.$$

We now state the convergence result for Algorithm 4.2.1. This result and later results make use of a quantity

$$F^* := \liminf_{k \to \infty} F_k(x_k), \tag{4.17}$$

which is finite if F is bounded below (and possibly in other circumstances too). When $F^* = -\infty$, the claim of the theorem is vacuously true.

Theorem 4.2.2. *Suppose that Algorithm 4.2.1 is applied to the problem* (4.1)*, where g is ρ-weakly convex and ∇h and g are Lipschitz continuous with constants $L_{\nabla h}$ and L_g, respectively. We have for all $k \geq 1$*

$$\begin{aligned}\min_{1\leq j\leq k} \; &\operatorname{dist}(-\nabla h(x_j), A^*\partial g(\operatorname{prox}_{\lambda_j g}(Ax_j))) \\ &\leq k^{-1/3}\sqrt{L_{\nabla h} + 2\rho\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1}L_g^2},\end{aligned}$$

where

$$\|Ax_j - \operatorname{prox}_{\lambda_j g}(Ax_j)\| \leq j^{-1/3}(2\rho)^{-1}L_g,$$

and F^* is defined as in (4.17)*. If A is also surjective, then for $x_k^* := x_k - A^\dagger(Ax_k - \operatorname{prox}_{\lambda_k g}(Ax_k))$, we have*

$$\begin{aligned}\min_{1\leq j\leq k} \; &\operatorname{dist}(-\nabla h(x_j^*), A^*\partial g(Ax_j^*)) \\ &\leq k^{-1/3}\left(\sqrt{L_{\nabla h} + (2\rho)\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1}L_g^2} + L_{\nabla h}\sigma_{\min}(A)^{-1}L_g\right)\end{aligned}$$

and $\|x_j - x_j^*\| \leq \sigma_{\min}(A)^{-1}L_g\lambda_j = \sigma_{\min}(A)^{-1}L_g(2\rho)^{-1}j^{-1/3}$.

Before proving this theorem, we state and prove a lemma that relates the function values of two Moreau envelopes with two different smoothing parameters. In the convex case, such statements are well known, but in the nonconvex case this result is novel.

Lemma 4.2.3. *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper, ρ-weakly convex and lower semicontinuous function, and let λ_2 and λ_1 be parameters such that $0 < \lambda_2 \leq \lambda_1 < \rho^{-1}$. Then, we have*

$$g_{\lambda_2}(y) \leq g_{\lambda_1}(y) + \frac{1}{2}\frac{\lambda_1 - \lambda_2}{\lambda_2}\lambda_1 \|\nabla g_{\lambda_1}(y)\|^2.$$

If, in addition, g is L_g-Lipschitz continuous, we have

$$g_{\lambda_2}(y) \leq g_{\lambda_1}(y) + \frac{1}{2}\frac{\lambda_1 - \lambda_2}{\lambda_2}\lambda_1 L_g^2.$$

Proof. By using the definition of the Moreau envelope, together with Lemma 2.3.7, we obtain

$$\begin{aligned}
g_{\lambda_2}(y) &= \min_{u \in \mathbb{R}^n} \left\{ g(u) + \frac{1}{2\lambda_2}\|y-u\|^2 \right\} \\
&= \min_{u \in \mathbb{R}^n} \left\{ g(u) + \frac{1}{2\lambda_1}\|y-u\|^2 + \frac{1}{2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\|y-u\|^2 \right\} \\
&\leq g(\operatorname{prox}_{\lambda_1 g}(y)) + \frac{1}{2\lambda_1}\|y - \operatorname{prox}_{\lambda_1 g}(y)\|^2 + \frac{1}{2}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right)\|y - \operatorname{prox}_{\lambda_1 g}(y)\|^2 \\
&= g_{\lambda_1}(y) + \frac{1}{2}\left(\frac{\lambda_1 - \lambda_2}{\lambda_2}\right)\lambda_1 \|\nabla g_{\lambda_1}(y)\|^2,
\end{aligned}$$

proving the first claim. The second claim follows immediately from (2.4). □

Proof of Theorem 4.2.2. Since $L_k = 1/\gamma_k$ is the Lipschitz constant of ∇F_k, we have for any $k \geq 1$ that

$$F_k(x_{k+1}) \leq F_k(x_k) + \langle \nabla F_k(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\gamma_k}\|x_{k+1} - x_k\|^2.$$

By substituting the definition of x_{k+1} from (4.16), we have

$$F_k(x_{k+1}) \leq F_k(x_k) - \frac{\gamma_k}{2}\|\nabla F_k(x_k)\|^2. \tag{4.18}$$

From Lemma 4.2.3, we have for all $x \in \mathbb{R}^d$

$$F_{k+1}(x) \leq F_k(x) + \frac{1}{2}(\lambda_k - \lambda_{k+1})\frac{\lambda_k}{\lambda_{k+1}}\|(\nabla g_{\lambda_k})(Ax)\|^2 \leq F_k(x) + (\lambda_k - \lambda_{k+1})L_g^2,$$

where we used in the second inequality that $\frac{\lambda_k}{\lambda_{k+1}} \leq 2$. We set $x = x_{k+1}$ and substitute into (4.18) to obtain

$$F_{k+1}(x_{k+1}) \leq F_k(x_k) - \frac{\gamma_k}{2}\|\nabla F_k(x_k)\|^2 + (\lambda_k - \lambda_{k+1})L_g^2.$$

By summing both sides of this expression over $k = 1, 2, \ldots, K$ for $K \geq 1$, and telescoping, we deduce

$$\sum_{k=1}^{K} \frac{\gamma_k}{2}\|\nabla F_k(x_k)\|^2 \leq F_1(x_1) - F_K(x_K) + (\lambda_1 - \lambda_K)L_g^2 \leq F_1(x_1) - F^* + \lambda_1 L_g^2. \tag{4.19}$$

Since

$$\gamma_k = \frac{\lambda_k}{\lambda_k L_{\nabla h} + \|A\|^2} \geq k^{-1/3} \frac{(2\rho)^{-1}}{(2\rho)^{-1} L_{\nabla h} + \|A\|^2} = k^{-1/3} \frac{1}{L_{\nabla h} + 2\rho\|A\|^2}$$

we have from (4.19) that

$$\frac{1}{L_{\nabla h} + 2\rho\|A\|^2} \min_{1\leq j\leq K} \|\nabla F_j(x_j)\|^2 \frac{1}{2} \sum_{k=1}^{K} k^{-1/3} \leq F_1(x_1) - F^* + (2\rho)^{-1} L_g^2. \qquad (4.20)$$

Now we observe that for all $K \geq 1$

$$\sum_{k=1}^{K} k^{-1/3} \geq \sum_{k=1}^{K} \int_k^{k+1} x^{-1/3}\,\mathrm{d}x = \int_1^{K+1} x^{-1/3}\,\mathrm{d}x = \frac{3}{2}\left((K+1)^{2/3} - 1\right)$$
$$\geq (K+1)^{2/3} - 1 \geq \frac{1}{2} K^{2/3},$$

where the final inequality can be checked numerically. Therefore, by substituting into (4.20), we have

$$\min_{1\leq j\leq K} \|\nabla F_j(x_j)\|^2 \leq 4 \frac{L_{\nabla h} + (2\rho)\|A\|^2}{K^{2/3}} \left(F_1(x_1) - F^* + (2\rho)^{-1} L_g^2\right)$$

and so

$$\min_{1\leq j\leq K} \|\nabla F_j(x_j)\| \leq \frac{C}{K^{1/3}},$$

where $C := 2\sqrt{L_{\nabla h} + (2\rho)\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1} L_g^2}$. By combining this bound with (4.8), and defining $z_j := \operatorname{prox}_{\lambda_j g}(Ax_j)$ for all $j = 1, \ldots, k$, we obtain

$$\min_{1\leq j\leq k} \operatorname{dist}(-\nabla h(x_j), A^* \partial g(z_j)) \leq \min_{1\leq j\leq k} \|\nabla F_j(x_j)\| \leq \frac{C}{k^{1/3}}, \qquad (4.21)$$

where we deduce from (2.5) that

$$\|Ax_j - z_j\| \leq \frac{(2\rho)^{-1} L_g}{j^{1/3}} \quad \forall j \geq 1.$$

The second statement concerning surjectivity of A follows from the consideration made in (4.10) to (4.12). □

There is a mismatch between the two bounds in this theorem. The first bound (the criticality bound) indicates that during the first $k = \mathcal{O}(\epsilon^{-3})$ iterations, we will encounter an iteration j at which the first-order optimality condition is satisfied within a tolerance of ϵ. However, this bound could have been satisfied at an early iteration (that is, $j \ll \epsilon^{-3}$), for which value the second (feasiblity) bound, on $\|Ax_j - \operatorname{prox}_{\lambda_j g}(Ax_j)\|$, may not be particularly small. The next section describes an algorithm that remedies this defect.

4.2.2 An epoch-wise approach

We describe a variant of Algorithm 4.2.1 in which the steps are organized into a series of epochs, each of which is twice as long as the one before. We show that there is some iteration $j = \mathcal{O}(\epsilon^{-3})$ such that both $\operatorname{dist}(-\nabla h(x_j), A^*\partial g(\operatorname{prox}_{\lambda_j g}(Ax_j)))$ and $\|Ax_j - \operatorname{prox}_{\lambda_j g}(Ax_j)\|$ are smaller than the given tolerance ϵ.

Algorithm 4.2.4 (Variable smoothing with epochs)**.**

```
Require: $x_1 \in \mathbb{R}^d$ and tolerance $\epsilon > 0$;
  for $l = 0, 1, \dots$ do
    Set $S_l \leftarrow \infty$, Set $j_l \leftarrow 2^l$;
    for $k = 2^l, 2^l + 1, \dots, 2^{l+1} - 1$ do
      Set $\lambda_k \leftarrow (2\rho)^{-1} k^{-1/3}$, define $L_k$ as in (4.15), set $\gamma_k \leftarrow 1/L_k$;
      Set $x_{k+1} \leftarrow x_k - \gamma_k \nabla F_k(x_k)$;
      if $\|\nabla F_{k+1}(x_{k+1})\| \leq S_l$ then
        Set $S_l \leftarrow \|\nabla F_{k+1}(x_{k+1})\|$; Set $j_l \leftarrow k + 1$;
        if $S_l \leq \epsilon$ and $\|Ax_{k+1} - \operatorname{prox}_{\lambda_{k+1} g}(Ax_{k+1})\| \leq \epsilon$ then
          STOP;
        end if
      end if
    end for
  end for
```

Theorem 4.2.5. *Consider solving* (4.1) *using Algorithm 4.2.4, where h and g satisfy the assumptions of Theorem 4.2.2 and F^* defined in* (4.17) *is finite. For a given tolerance $\epsilon > 0$, Algorithm 4.2.4 generates an iterate x_j for some $j = O(\epsilon^{-3})$ such that*

$$\operatorname{dist}(-\nabla h(x_j), A^*\partial g(z_j)) \leq \epsilon \quad \textit{and} \quad \|Ax_j - z_j\| \leq \epsilon, \quad \textit{where } z_j = \operatorname{prox}_{\lambda_j g}(Ax_j).$$

Proof. As in (4.19), by using monotonicity of $(F_k(x_k))_{k\geq 1}$ and discarding nonnegative terms, we have that for all $l \geq 1$

$$\sum_{k=2^l}^{2^{l+1}-1} \frac{\gamma_k}{2} \|\nabla F_k(x_k)\|^2 \leq F_1(x_1) - F^* + (2\rho)^{-1} L_g^2.$$

With the same arguments as in the earlier proof, we obtain

$$\begin{aligned}\sum_{k=2^l}^{2^{l+1}-1} k^{-1/3} &\geq \sum_{k=2^l}^{2^{l+1}-1} \int_k^{k+1} x^{-1/3}\,\mathrm{d}x = \int_{2^l}^{2^{l+1}} x^{-1/3}\,\mathrm{d}x = \frac{3}{2}\left((2^{l+1})^{2/3} - (2^l)^{2/3}\right) \\ &= \frac{3}{2}\left(2^{2/3} - 1\right)(2^l)^{2/3} \geq \frac{1}{2}(2^l)^{2/3}.\end{aligned}$$

Therefore, we have

$$\min_{2^l \leq j \leq 2^{l+1}-1} \|\nabla F_j(x_j)\| \leq \frac{C}{(2^l)^{1/3}},$$

with C defined as before, that is, $C = 2\sqrt{L_{\nabla h} + (2\rho)\|A\|^2}\sqrt{F_1(x_1) - F^* + (2\rho)^{-1}L_g^2}$. Noting that $z_j := \operatorname{prox}_{\lambda_j g}(Ax_j)$, we have as in (4.21) that

$$\min_{2^l \leq j \leq 2^{l+1}-1} \operatorname{dist}(-\nabla h(x_j), A^*\partial g(z_j)) \leq \frac{C}{(2^l)^{1/3}}, \tag{4.22}$$

as previously. Further, we have for $2^l \leq j \leq 2^{l+1} - 1$ that

$$\|Ax_j - z_j\| \leq L_g \lambda \leq \frac{(2\rho)^{-1}L_g}{j^{1/3}} \leq \frac{(2\rho)^{-1}L_g}{(2^l)^{1/3}}. \tag{4.23}$$

From (4.22) and (4.23) we deduce that Algorithm 4.2.4 must terminate before the end of epoch l, that is, before 2^{l+1} iterations have been completed, where l is the first nonnegative integer such that

$$2^l \geq \max\{C^3, (2\rho)^{-3}L_g^3\}\epsilon^{-3}.$$

Thus, termination occurs after at most $2\max\{C^3, (2\rho)^{-3}L_g^3\}\epsilon^{-3}$ iterations. □

For the case of A surjective, we have the following stronger result.

Corollary 4.2.6. *Suppose that the assumptions of Theorem 4.2.5 hold, that A is also surjective, and that the condition $\|Ax_{k+1} - \operatorname{prox}_{\lambda_{k+1}g}(Ax_{k+1})\| \leq \epsilon$ in Algorithm 4.2.4 is replaced by $\|x_{k+1} - x_{k+1}^*\| \leq \epsilon$, for $x_j^* := x_j - A^\dagger(Ax_j - \operatorname{prox}_{\lambda_j g}(Ax_j))$. Then for some $j' = O(\epsilon^{-3})$, we have that*

$$\operatorname{dist}(-\nabla h(x_{j'}^*), A^*\partial g(Ax_{j'}^*)) \leq \epsilon$$

and $\|x_{j'} - x_{j'}^*\| \leq \epsilon$.

Proof. With the considerations made in the previous proof as well as the one made in (4.10) to (4.12), we can choose l to be the smallest positive integer such that

$$2^{l+1} \geq 2\max\{C^3, \sigma_{\min}(A)^{-3}L_g^3(2\rho)^{-3}\}\epsilon^{-3}.$$

The claim then holds for some $j' \leq 2^{l+1}$. □

Although Algorithm 4.2.4 seems more complicated than Algorithm 4.2.1, the steps are the same. The only difference is that for the second algorithm, we do not search for the iterate that minimizes criticality across *all* iterations but only across at most the last $k/2$ iterations, where k is the total number of iterations.

4.3 Proximal gradient

Here we derive a complexity bound for the proximal gradient algorithm applied to the more elementary problem (4.4) studied in Section 4.1.3, that is,

$$\min_{x\in\mathbb{R}^d} F(x) := h(x) + g(x), \tag{4.24}$$

for $h : \mathbb{R}^d \to \mathbb{R}$ a $L_{\nabla h}$-smooth function and $g : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ a possibly nonsmooth, but proper, ρ-weakly convex and lower semicontinuous function. Such a bound has not been made explicit before, to the authors' knowledge, though it is a fairly straightforward consequence of existing results. The bound makes a interesting comparison with the result in Section 4.2, where the nonsmoothness issue becomes more complicated due to the composition with a linear operator. In this section, we assume that a closed-form proximal operator is available for g, and we show that the complexity bound of $\mathcal{O}(\epsilon^{-2})$ is the same order as for gradient descent applied to smooth nonconvex functions.

Standard proximal gradient for (4.24), given parameter $\gamma \in (0, \min\{\rho^{-1}/2, L_{\nabla h}^{-1}\}]$ and initial point x_1, is as follows:

$$\begin{aligned} x_{k+1} &:= \arg\min_{x\in\mathbb{R}^d} \left\{ g(x) + \langle \nabla h(x_k), x - x_k\rangle + \frac{1}{2\gamma}\|x - x_k\|^2 \right\}, \qquad (4.25)\\ &= \operatorname{prox}_{\gamma g}\left(x_k - \gamma\nabla h(x_k)\right), \quad k = 1, 2, \ldots, \end{aligned}$$

where the choice of γ ensures that the function to be minimized in (4.25) is $(\gamma^{-1} - \rho)$-strongly convex, so that x_{k+1} is uniquely defined.

We have the following convergence result.

Theorem 4.3.1. *Consider the algorithm defined by* (4.25) *applied to problem* (4.24), *where we assume that g is ρ-weakly convex and that ∇h is Lipschitz continuous with constant $L_{\nabla h}$. Supposing that $\gamma \in (0, \min\{\rho^{-1}/2, L_{\nabla h}^{-1}\}]$, we have for all $k \geq 1$ that*

$$\min_{2\leq j\leq k+1} \operatorname{dist}(0, \partial(h+g)(x_j)) \leq k^{-1/2}\sqrt{2(F(x_1) - F^*)}\,\frac{\gamma^{-1} + L_{\nabla h}}{\sqrt{\gamma^{-1} - \rho}},$$

where F^* is defined in (4.17).

Proof. Note first that the result is vacuous if $F^* = -\infty$, so we assume henceforth that F^* is finite. We have for every $x \in \mathbb{R}^d$ that

$$\begin{aligned} g(x_{k+1}) + h(x_k) + \langle \nabla h(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 + \frac{1}{2}(\gamma^{-1} - \rho)\|x - x_{k+1}\|^2 \\ \leq g(x) + h(x_k) + \langle \nabla h(x_k), x - x_k\rangle + \frac{1}{2\gamma}\|x - x_k\|^2. \end{aligned}$$

By applying the inequality

$$h(x_{k+1}) \leq h(x_k) + \langle \nabla h(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\gamma}\|x_{k+1} - x_k\|^2 \quad \text{for all } x \in \mathbb{R}^d,$$

obtained from the Lipschitz continuity of ∇h and the fact that $\gamma \leq L_{\nabla h}^{-1}$, we deduce that

$$F(x_{k+1}) + \frac{1}{2}(\gamma^{-1} - \rho)\|x - x_{k+1}\|^2 \leq g(x) + h(x_k) + \langle \nabla h(x_k), x - x_k \rangle + \frac{1}{2\gamma}\|x - x_k\|^2,$$

for every $x \in \mathbb{R}^d$. By setting $x = x_k$, we obtain

$$F(x_{k+1}) + \frac{1}{2}(\gamma^{-1} - \rho)\|x_k - x_{k+1}\|^2 \leq F(x_k),$$

which shows, together with the definition (4.17), that

$$\sum_{k=1}^{\infty} \|x_k - x_{k+1}\|^2 \leq \frac{2(F(x_1) - F^*)}{\gamma^{-1} - \rho}. \tag{4.26}$$

From the optimality conditions for (4.25), we obtain

$$0 \in \nabla h(x_k) + \partial g(x_{k+1}) + \gamma^{-1}(x_{k+1} - x_k)$$

which also shows that

$$w_{k+1} := \frac{1}{\gamma}(x_k - x_{k+1}) + \nabla h(x_{k+1}) - \nabla h(x_k) \in \partial(h + g)(x_{k+1}), \tag{4.27}$$

so that

$$\|w_{k+1}\|^2 \leq (\gamma^{-1} + L_{\nabla h})^2 \|x_k - x_{k+1}\|^2.$$

By combining this bound with (4.26), we obtain

$$\sum_{k=1}^{\infty} \|w_{k+1}\|^2 \leq 2(F(x_1) - F^*)\frac{(\gamma^{-1} + L_{\nabla h})^2}{\gamma^{-1} - \rho},$$

from which it follows that

$$\min_{1 \leq j \leq k} \|w_{j+1}\| \leq \sqrt{2(F(x_1) - F^*)}\,\frac{(\gamma^{-1} + L_{\nabla h})}{\sqrt{k}\sqrt{\gamma^{-1} - \rho}}.$$

The result now follows from (4.27), when we note that

$$\min_{1 \leq j \leq k} \operatorname{dist}(0, \partial(h + g)(x_{j+1})) \leq \min_{1 \leq j \leq k} \|w_{j+1}\|.$$

□

This theorem indicates that the proximal gradient algorithm requires at most $\mathcal{O}(\epsilon^{-2})$ to find an iterate with ϵ-approximate stationarity. This bound contrasts with the bound $\mathcal{O}(\epsilon^{-3})$ of Section 4.2 for the case of general A. Moreover, the $\mathcal{O}(\epsilon^{-2})$ bound has the same order as the bound for gradient descent applied to general smooth nonconvex optimization.

5 Convex-concave minimax problems

Motivated by the training of Generative Adversarial Networks (GANs), we study methods for solving minimax problems with additional nonsmooth regularizers. We do so by employing *monotone operator* theory, in particular the *Forward-Backward-Forward (FBF)* method, which avoids the known issue of limit cycling by correcting each update by a second gradient evaluation. Furthermore, we propose a seemingly new scheme which recycles old gradients to mitigate the additional computational cost. In doing so we rediscover a known method, related to *Optimistic Gradient Descent Ascent (OGDA)*. For both schemes we prove novel convergence rates for convex-concave minimax problems via a unifying approach. The derived error bounds are in terms of the gap function for the ergodic iterates. For the deterministic and the stochastic problem we show a convergence rate of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$, respectively. We complement our theoretical results with empirical improvements in the training of Wasserstein GANs on the CIFAR10 dataset.

5.1 About GANs

Generative Adversarial Networks (GANs) [43] have proven to be a powerful class of generative models, producing for example unseen realistic images. Two neural networks, called generator and discriminator, compete against each other in a game. In the special case of a zero sum game this task can be formulated as a minimax problem.

Conventionally, GANs are trained using variants of (stochastic) *Gradient Descent Ascent (GDA)* which are known to exhibit oscillatory behavior and thus fail to converge even for simple bilinear saddle point problems, see [42]. We therefore propose the use of methods with provable convergence guarantees for (stochastic) convex-concave minimax problems, even though GANs are well known to not warrant these properties. Along similar considerations an adaptation of the *Extragradient method (EG)* [54] for the training of GANs was suggested in [40], whereas [33, 34, 58] studied *Optimistic Gradient Descent Ascent (OGDA)* based on *optimistic mirror descent* [87, 88]. We however investigate the *Forward-Backward-Forward (FBF)* method [103] from monotone operator theory, which uses two gradient evaluations per update, similar to EG, in order to circumvent the aforementioned issues.

Instead of trying to improve GAN performance via new architectures, loss functions, etc., we contribute to the theoretical foundation of their training from the point of view of optimization.

Contribution. Establishing the connection between GAN training and *monotone inclusions* [6] motivates to use the FBF method, originally designed to solve this type of problems. This approach allows to naturally extend the constrained setting to a regularized one making use of the proximal operator.

We also propose a variant of FBF reusing previous gradients to reduce the computational cost per iteration, which turns out to be a known method, related to OGDA. By developing a unifying scheme that captures FBF and a generalization of OGDA, we reveal a hitherto unknown connection. Using this approach we prove novel non asymptotic convergence statements in terms of the minimax gap for both methods in the context of saddle point problems. In the deterministic and stochastic setting we obtain rates of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$, respectively. Concluding, we highlight the relevance of our proposed method as well as the role of regularizers by showing empirical improvements in the training of Wasserstein GANs on the CIFAR10 dataset.

Organization. In Section 5.2 we highlight the connection of GAN training and monotone inclusions and give an extensive review of methods with convergence guarantees for the latter. The main results as well as a precise definition of the measure of optimality are discussed in Section 5.3. Concluding, Section 5.4 illustrates the empirical performance in the training of GANs as well as solving bilinear problems.

5.2 GAN training as monotone inclusion

The GAN objective was originally cast as a two-player zero-sum game (see [43]) between the discriminator D_y and the generator G_x given by

$$\min_x \max_y \mathbb{E}_{\rho\sim q}[\log(D_y(\rho))] + \mathbb{E}_{\zeta\sim p}[\log(1 - D_y(G_x(\zeta)))],$$

exhibiting the aforementioned minimax structure. Due to problems with vanishing gradients in the training of such models, a successful alternative formulation called *Wasserstein GAN (WGAN)* [2] has been proposed. In this case the minimization tries to reduce the Wasserstein distance between the true distribution q and the one learned by the generator. Reformulating this distance via the Kantorovich-Rubinstein duality leads to an inner maximization over 1-Lipschitz functions which are approximated via neural networks, yielding the saddle point problem

$$\min_x \max_y \mathbb{E}_{\rho\sim q}[D_y(\rho)] - \mathbb{E}_{\zeta\sim p}[D_y(G_x(\zeta))].$$

5.2.1 Convex-concave minimax problems

Due to the observations made in the previous paragraph we study the following abstract minimax problem

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^n} \Psi(x,y) := f(x) + \mathbb{E}_{\xi\sim Q}\left[\Phi(x,y;\xi)\right] - h(y), \tag{5.1}$$

where the convex-concave coupling function $\Phi(x,y) := \mathbb{E}_{\xi\sim Q}\left[\Phi(x,y;\xi)\right]$ is differentiable with L-Lipschitz continuous gradient. The proper, convex and lower semicontinuous functions $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ act as regularizers. A solution of (5.1) is given by a so-called *saddle point* (x^*, y^*) fulfilling for all x and y

$$\Psi(x^*, y) \leq \Psi(x^*, y^*) \leq \Psi(x, y^*).$$

In the context of two-player games this corresponds to a pair of strategies, where no player can be better off by changing just their own strategy.

For illustrative purposes, we will restrict ourselves for now to the special case of the *deterministic constrained* version of (5.1), given by

$$\min_{x \in X} \max_{y \in Y} \Phi(x, y)$$

where f and h are given by indicator functions of nonempty, convex and closed sets X and Y, respectively.

5.2.2 Minimax problems as monotone inclusions

If the coupling function Φ is convex-concave and differentiable then the necessary and sufficient optimality condition can be written as a so-called *monotone inclusion* using

$$F(x, y) := (\nabla_x \Phi(x, y), -\nabla_y \Phi(x, y)) \tag{5.2}$$

and the *normal cone* N_Ω of the convex set $\Omega := X \times Y$. By denoting $w = (x, y) \in \mathbb{R}^m$ where $m = d + n$, it reads

$$0 \in F(w) + N_\Omega(w). \tag{5.3}$$

The normal cone mapping is given by

$$N_\Omega(w) = \{v \in \mathbb{R}^m : \langle v, w' - w \rangle \leq 0 \quad \forall w' \in \Omega\},$$

for $w \in \Omega$ and $N_\Omega(w) = \emptyset$ for $w \notin \Omega$. Here, the operators F and N_Ω satisfy well known properties from convex analysis [6], in particular the first one is monotone (and Lipschitz if $\nabla\Phi$ is so) whereas the latter one is maximal monotone. We call a, possibly *set-valued*, operator A from $\mathbb{R}^m$ to itself monotone [6] if

$$\langle u - u', z - z' \rangle \geq 0 \quad \forall u \in A(z), u' \in A(z').$$

We say A is maximal monotone, if there exists no monotone operator A' such that the graph of A is properly contained in the graph of A'.

Problems of type (5.3) have been studied thoroughly in convex optimization, with the most established solution methods being *Extragradient (aka Korpelevich)* [54] and *Forward-Backward-Forward (aka Tseng)* [103]. Both methods are known to generate sequences of iterates converging to a solution of (5.3). Note that in the unconstrained setting (i.e. if Ω is the entire space) both of these algorithms even produce the same iterates.

5.2.3 Solving monotone inclusions

The connection between monotone inclusions and saddle point problems is of course not new. The application of Extragradient (EG) to minimax problems has been studied in the seminal paper [71] under the name of *Mirror Prox* and a convergence rate of $\mathcal{O}(1/k)$ in terms of the function values has been proven. Even a stochastic version of the Mirror Prox algorithm has been studied in [51] with a convergence rate of $\mathcal{O}(1/\sqrt{k})$. Applied to problem (5.3), with P_Ω being the projection onto Ω, it iterates

$$\text{EG:} \left\lfloor \begin{array}{l} w_k = P_\Omega[z_k - \alpha_k F(z_k)] \\ z_{k+1} = P_\Omega[z_k - \alpha_k F(w_k)]. \end{array} \right.$$

The Forward-Backward-Forward (FBF) method has not been studied rigorously for minimax problems yet, despite promising applications in [24] and its advantage of it only requiring one projection, whereas EG needs two. It is given by

$$\text{FBF:} \left\lfloor \begin{array}{l} w_k = P_\Omega[z_k - \alpha_k F(z_k)] \\ z_{k+1} = w_k + \alpha_k (F(z_k) - F(w_k)). \end{array} \right. \tag{5.4}$$

Both, EG and FBF, have the "disadvantage" of needing two gradient evaluations per iteration. A possible remedy — suggested in [40] for EG under the name of *extrapolation from the past* — is to recycle previous gradients. In a similar fashion we introduce

$$\text{FBFp:} \left\lfloor \begin{array}{l} w_k = P_\Omega[z_k - \alpha_k F(w_{k-1})] \\ z_{k+1} = w_k + \alpha_k (F(w_{k-1}) - F(w_k)), \end{array} \right. \tag{5.5}$$

where we replaced $F(z_k)$ by $F(w_{k-1})$ twice in (5.4). As a matter of fact, the above method can be written exclusively in terms of the first variable w_k by incrementing the index k in the first update and then substituting in the second line. This results in

$$w_{k+1} = P_\Omega\Big[w_k - \alpha_{k+1} F(w_k) + \alpha_k (F(w_{k-1}) - F(w_k))\Big]. \tag{5.6}$$

This way we rediscover a known method which was studied in [66] for general monotone inclusions under the name of *forward-reflected-backward*. It reduces to *optimistic mirror descent* [87, 88] in the unconstrained case with constant stepsize $\alpha_k = \alpha$, giving

$$w_{k+1} = w_k - \alpha(2F(w_k) - F(w_{k-1})) \tag{5.7}$$

which has been proposed for the training of GANs under the name of *Optimistic Gradient Descent Ascent (OGDA)*, see [33, 34, 58].

All of the above methods and extensions rely solely on the monotone operator formulation of the saddle point problem where the two components x and y play a symmetric role. Taking the special minimax structure into consideration, [46] showed convergence of a method that uses an optimistic step (5.7) in one component and a regular gradient step in the other, thus requiring less storing of past gradients in comparison to (5.6).

On the downside, however, by reducing the number of required gradient evaluations per iteration, the largest possible stepsize is reduced from $1/L$ (see [54] or Section 5.3) to $1/2L$ (see [40, 65, 66] or Section 5.3). To summarize, the number of required gradient evaluations is halved, but so is the stepsize, resulting in no clear net gain.

5.2.4 Regularizers

The role of regularizers is well studied in many fields such as statistics [100], signal processing [81] or inverse problems [94]. They serve different purposes such as inducing sparsity in the solution or conditioning of the problem. In the context of deep learning this has been explored from different perspectives, e.g. in incremental convex neural networks where neurons with zero weights are removed from the network and new ones are inserted according to different policies, see [3, 10, 84, 93].

In the framework of monotone operator theory the optimality condition of the regularized minimax problem (5.1) can be written as

$$0 \in F(w) + \partial r(w), \tag{5.8}$$

where r is given by $(x, y) \mapsto f(x) + h(y)$ and ∂r denotes its subdifferential, see Definition 2.1.7. The monotone inclusion (5.8) generalizes (5.3) in a natural way, since $N_\Omega = \partial \delta_\Omega$. In particular, the proximal mapping of the indicator δ_Ω yields the projection onto the set Ω, i.e. $\operatorname{prox}_{\lambda\delta_\Omega} = P_\Omega$.

5.3 Main results

Motivated by the considerations above we study the inclusion problem

$$0 \in F(w) + \partial r(w), \tag{5.9}$$

where $F : \mathbb{R}^m \to \mathbb{R}^m$ is a monotone and Lipschitz operator and $r : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex and lower semicontinuous function.

5.3.1 Measure of optimality

Typically in monotone inclusions, the distance to the set of solutions is used as a measure of quality of a given point due to the lack of more specific structure in general. Asymptotic convergence of the iterates has been established for FBF and FBFp in [6, Proposition 27.13] and [66], respectively. Furthermore, no convergence rates can be expected without stronger monotonicity assumptions. We will therefore focus on the following *gap function*, given for any $w \in \mathbb{R}^m$ by

$$\sup_{z \in \mathbb{R}^m} \langle F(z), w - z \rangle + r(w) - r(z),$$

for which we will be able to prove quantitative convergence rates. If r is the indicator δ_Ω of the compact and convex set Ω it is clear that the supremum is only taken over $z \in \Omega$ and will thus be finite. Since the problem (5.9) is in general unconstrained and the supremum can be infinite we consider instead, as done in e.g. [75], the *restricted* gap where the above supremum is taken over an auxiliary compact set $B \subset \mathbb{R}^m$ instead of the entire space.

$$G_B^{VI}(w) = \sup_{z \in B} \langle F(z), w - z \rangle + r(w) - r(z), \tag{5.10}$$

where we interpret the possible occurrence of $\infty - \infty$ as $+\infty$. Note that the restricted gap is in general only a reasonable measure of optimality for elements of B.

If F arises from a saddle point problem (5.1) meaning that F has the form (5.2), we want to use a more problem specific measure, the minimax gap, which for a point $w = (u, v) \in \mathbb{R}^d \times \mathbb{R}^n$ is given by

$$G_B^{SP} = \sup_{(x,y)\in B} \Psi(u, y) - \Psi(x, v). \tag{5.11}$$

In order to capture both at the same time we define the following unifying gap

$$G_B(w) := \begin{cases} \sup_{(x,y)\in B} \Psi(u, y) - \Psi(x, v) & \text{if } F \text{ and } r \text{ come from (5.1)} \\ \sup_{z\in B} \langle F(z), w - z\rangle + r(w) - r(z) & \text{otherwise.} \end{cases} \tag{5.12}$$

use the following (restricted) *minimax gap*, common for saddle point problems, which for a point (u, v) is given by

$$G_B(u, v) = \sup_{(x,y)\in B} \Psi(u, y) - \Psi(x, v).$$

For the general case, i.e. F being an arbitrary monotone and Lipschitz operator this is connected to the other measure of optimality we use in (5.12), for $w \in \mathbb{R}^m$ given by It stems from the field of Variational Inequalities where such a function is also known as *merit function* [75]. The relevance of the above two quantities will be made clear by the following statements.

Theorem 5.3.1. *Let $\Phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, convex and lower semicontinuous and $B \subset \mathbb{R}^d \times \mathbb{R}^n$. A point (x^*, y^*) in the interior of B solves the saddle point problem* (5.1) *if and only if its minimax gap* (5.11) *is zero, $G_B^{SP}(x^*, y^*) = 0$. For all other elements of B the gap is nonnegative.*

Proof. A saddle point (x^*, y^*) clearly fulfills that $\sup_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^n} \Psi(x^*, y) - \Psi(x, y^*) = 0$. On the other hand let $G_B^{SP}(x^*, y^*) = 0$. For an arbitrary point (x, y) we can choose $\alpha \in (0, 1)$ large enough such that $(u, v) := \alpha(x^*, y^*) + (1 - \alpha)(x, y)$ is in the interior of B. Therefore,

$$\Psi(x^*, v) - \Psi(u, y^*) = \Psi(x^*, \alpha y^* + (1 - \alpha)y) - \Psi(\alpha x^* + (1 - \alpha)x, y^*) \leq 0.$$

Using the convex-concave structure of Ψ we deduce that

$$\alpha\Psi(x^*, y^*) + (1 - \alpha)\Psi(x^*, y) - \alpha\Psi(x^*, y^*) - (1 - \alpha)\Psi(x, y^*) \leq 0,$$

which implies that $\Psi(x^*, y) \leq \Psi(x, y^*)$. Since (x, y) was chosen arbitrary (x^*, y^*) is a saddle point. □

Similarly, an analogous statement can be shown for (5.10). The proof, however is split up into multiple lemmas to highlight the connection to Variational Inequalities.

Theorem 5.3.2. *Let $F : \mathbb{R}^m \to \mathbb{R}^m$ be monotone and continuous, $r : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ proper, convex and lower semicontinuous and $B \subset \mathbb{R}^m$. A point w^* in the interior of B solves the monotone inclusion*

$$0 \in F(w) + \partial r(w) \tag{5.13}$$

if and only if its restricted gap (5.10) *is zero, $G_B^{VI}(w^*) = 0$. For all other elements of B the gap is nonnegative.*

Let the assumptions of Theorem 5.3.2 hold true for the following lemmas as we break up the proof into separate statements. We do so by making use of the associated *Variational inequality (VI)*

$$\text{find } w \text{ such that } \quad \langle F(w), z - w \rangle + r(z) - r(w) \geq 0 \quad \forall z \in \mathbb{R}^m. \tag{5.14}$$

Lemma 5.3.3. *The monotone inclusion* (5.13) *is equivalent to the VI* (5.14)*.*

Proof. The equivalence of (5.13) and (5.14) follows immediately from the definition of the subdifferential of r. □

The formulation (5.14) is typically referred to as the *strong* form of the VI, whereas

$$\text{find } w \text{ such that } \quad \langle F(z), z - w \rangle + r(z) - r(w) \geq 0 \quad \forall z \in \mathbb{R}^m, \tag{5.15}$$

is known as the *weak* formulation.

Lemma 5.3.4. *Under the given assumptions the notion of weak and strong VI are equivalent.*

Proof. For the monotone operator F it is clear that if w^* is a solution to the strong formulation (5.14), it is also a solution to the weak formulation (5.15). In fact, if F is continuous the reverse implication also holds true. To see this, let w^* be a solution to the weak VI (5.15) and $z = \alpha w^* + (1-\alpha)u$ for an arbitrary $u \in \mathbb{R}^m$ and $\alpha \in (0,1)$, then

$$\langle F(\alpha w^* + (1-\alpha)u), (1-\alpha)(u - w^*) \rangle + r(\alpha w^* + (1-\alpha)u) - r(w^*) \geq 0.$$

This implies by the convexity of r that

$$(1-\alpha)\langle F(\alpha w^* + (1-\alpha)u), (u - w^*) \rangle + (1-\alpha)(r(u) - r(w^*)) \geq 0.$$

By dividing by $(1-\alpha)$ and then taking the limit $\alpha \to 1$ we obtain that w^* is a solution of the strong form (5.14). □

With the notion of VIs in mind, the above defined gap (5.10) becomes natural as it measures how much the statement of (5.15) is violated.

Lemma 5.3.5. *G_B^{VI} is nonnegative on B and zero for solutions of the weak VI.*

Proof. It is clear that $G_B^{VI}(w) \geq 0$ for $w \in B$ as $z = w$ can be chosen in the supremum. On the other hand if $w^* \in B$ is a solution to the weak VI (5.15) then $G_B^{VI}(w^*) = 0$. This follows from the fact that for a solution of (5.15) for all $z \in B$

$$\langle F(z), w^* - z\rangle + r(w^*) - r(z) \leq 0.$$

Therefore the supremum over the above expression in z is also less than zero, but clearly zero is obtained for $z = w^*$. □

For the reverse implication to hold true, we may not use points on the boundary of B.

Lemma 5.3.6. *If a point w^* in the interior of B exhibits zero gap $G_B^{VI}(w^*) = 0$, then it is a solution to the weak VI* (5.15).

Proof. Since w^* is in the interior of B we can, for an arbitrary $w \in \mathbb{R}^m$, choose $\alpha \in (0, 1)$ large enough such that $z := \alpha w^* + (1 - \alpha)w \in B$. Using this z in the supremum of the gap we deduce that

$$\langle F(\alpha w^* + (1 - \alpha)w), w^* - \alpha w^* - (1 - \alpha)w\rangle + r(w^*) - r(\alpha w^* + (1 - \alpha)w) \leq 0.$$

This implies that

$$(1 - \alpha)\langle F(\alpha w^* + (1 - \alpha)w), w - w^*\rangle + (1 - \alpha)(r(w) - r(w^*)) \geq 0.$$

By dividing by $(1 - \alpha)$ and then taking the limit $\alpha \to 1$ we deduce that w^* solves the strong form of the VI (5.14). □

Now, we can turn to proving the theorem.

Proof of Theorem 5.3.2. Combine Lemma 5.3.3, 5.3.4, 5.3.5 and 5.3.6. □

5.3.2 Methods

We now present a novel unifying scheme for solving problem (5.9), which generalizes FBF (5.4) and in addition recovers the method motivated in (5.5) as FBFp. Let us point out again that the latter algorithm was already introduced in [66] and corresponds to OGDA [33, 34, 87] if F stems from the minimax setting (5.2).

Algorithm 5.3.7 (generalized FBF)**.** For a starting point $z_0 \in \mathbb{R}^m$ and stepsizes $\alpha_k > 0$ we consider for all $k \geq 0$

$$\left\lfloor \begin{array}{l} w_k = \operatorname{prox}_{\alpha_k r}(z_k - \alpha_k F(\diamondsuit_k)) \\ z_{k+1} = w_k + \alpha_k (F(\diamondsuit_k) - F(w_k)). \end{array} \right.$$

For $\diamondsuit_k = z_k$ this reduces to the well known FBF method, whereas $\diamondsuit_k = w_{k-1}$, with the additional initial condition $w_{-1} = z_0$, recycles previous gradients (FBFp).

Consider the scenario where F is given as an expectation $\mathbb{E}_\xi[F(\cdot\,;\xi)]$, e.g. coming from (5.1), and only a stochastic estimator $F(\cdot\,;\xi)$ is accessible instead of F itself. In this case we adapt Algorithm 5.3.7 in the following way.

Algorithm 5.3.8 (generalized stochastic FBF)**.** For a starting point $z_0 \in \mathbb{R}^m$ and stepsizes $\alpha_k > 0$ we consider for all $k \geq 0$

$$\left\lfloor \begin{array}{l} \xi_k \sim Q \quad \text{(optionally } \eta_k \sim Q\text{)} \\ w_k = \operatorname{prox}_{\alpha_k r}\left(z_k - \alpha_k F(\diamondsuit_k; \triangle_k)\right) \\ z_{k+1} = w_k + \alpha_k (F(\diamondsuit_k; \triangle_k) - F(w_k; \xi_k)). \end{array} \right.$$

For $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$ this results in a stochastic version of FBF, whereas $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$ recycles previous gradients (stochastic FBFp) with the additional initial condition $w_{-1} = z_0$ and $\xi_{-1} = \eta_0$.

Even though both methods encompassed by the unifying scheme Algorithm 5.3.7 have been studied in the deterministic setting before, the stated convergence results are new. However, we want to point out that the stochastic version of FBFp has not been considered prior to this work.

5.3.3 Convergence

Let in the following $B \subset \mathbb{R}^m$ be the compact set of the restricted (unifying) gap function (5.12) with $D := \sup_{w,z \in B} \|z - w\|$ denoting its diameter. For convenience in the estimation we assume that the starting point z_0 of the discussed methods is in B. Lastly, all the convergence statement will be in terms of the *averaged iterates*, given for $K \geq 1$ by

$$\bar{w}_K := \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \sum_{k=0}^{K-1} \alpha_k w_k.$$

Theorem 5.3.9 (deterministic)**.** *Let $(w_k)_{k\geq 0}$ be the sequence generated by Algorithm 5.3.7. If*

(i) FBF, i.e. $\diamondsuit_k = z_k$, with stepsize $0 < \alpha_k \leq 1/L$, or

(ii) FBFp, i.e. $\diamondsuit_k = w_{k-1}$, with stepsize $0 < \alpha_k \leq 1/2L$

is chosen, then for all $K \geq 1$ the averaged iterates $\bar{w}_K := \frac{1}{K}\sum_{k=0}^{K-1} w_k$ fulfill

$$G_B(\bar{w}_K) \leq \frac{D^2}{2\sum_{k=0}^{K-1} \alpha_k}.$$

where G_B is the restricted gap defined in (5.12)*. For constant stepsize this results in a convergence rate of $\mathcal{O}(1/k)$.*

In order to derive similar convergence statements for the stochastic algorithm we need to assume (standard) properties of the gradient estimator $F(\cdot\,;\xi)$.

Assumption 5.3.10. *Unbiasedness:* $\mathbb{E}_\xi[F(w;\xi)] = F(w)\,\forall w \in \mathbb{R}^m$.

Assumption 5.3.11. *Bounded variance:* $\mathbb{E}_\xi[\|F(w;\xi) - F(w)\|^2] \leq \sigma^2\,\forall w \in \mathbb{R}^m$.

In particular we actually only need the above assumption to hold for all iterates w_k. Such an hypothesis is in practice difficult to check, but could be exploited in special cases where additional properties of the variance and boundedness of the iterates are known a priori.

Assumption 5.3.12. *The samples ξ_k are independent of the iterates w_k, for all $k \geq 0$.*

Equipped with these assumptions we are now able to prove the statement.

Theorem 5.3.13 (stochastic)**.** *Let Assumption 5.3.10, 5.3.11 and 5.3.12 hold and let $(w_k)_{k\geq 0}$ be the sequence generated by Algorithm 5.3.8. Then, with G_B being the restricted gap defined in* (5.12)*:*

(i) If stochastic FBF, i.e. $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$, with stepsize $\alpha_k \leq \alpha < \frac{1}{L}$ is chosen, then

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 4(1-\alpha^2L^2)^{-1}\sigma^2\sum_{k=0}^{K-1}\alpha_k^2}{2\sum_{k=0}^{K-1}\alpha_k}.$$

(ii) If stochastic FBFp, i.e. $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$, with stepsize $\alpha_k \leq \alpha < \frac{1}{2\sqrt{2}L}$, is chosen, then

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 2\left(5 + \frac{4\alpha^2L^2}{1-8\alpha^2L^2}\right)\sigma^2\sum_{k=0}^{K-1}\alpha_k^2}{2\sum_{k=0}^{K-1}\alpha_k}.$$

Although, the stepsize in the above statements of Theorem 5.3.13 can be chosen arbitrarily close to $1/L$ and $1/(2\sqrt{2}L)$ for stochastic FBF and stochastic FBFp, respectively. This does not mean it should be — since the constant in the convergence rate deteriorates when the stepsize is close to its allowed upper bound. The constants in the convergence rate for stochastic FBF(p) can, however, be combined and simplified by restricting the upper bound for the stepsizes α further. If $\alpha \leq 1/\sqrt{2}L$ for FBF, or $\alpha \leq 1/3L$ for FBFp, then

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 18\sigma^2\sum_{k=0}^{K-1}\alpha_k^2}{2\sum_{k=0}^{K-1}\alpha_k},$$

This statement exhibits a classical stepsize dependence [89], yielding convergence for sequences $(\alpha_k)_{k\geq 0}$ that are square summable $\sum_{k=0}^{\infty}\alpha_k^2 < +\infty$ but not summable $\sum_{k=0}^{\infty}\alpha_k = +\infty$. Additionally, if the stepsize is chosen $\alpha_k = \alpha/\sqrt{k+1}$, a convergence rate can be obtained and is given by

$$\mathbb{E}[G_B(\bar{w}_K)] = \mathcal{O}\Big(\frac{1}{\sqrt{K}}\Big). \tag{5.16}$$

If the stepsize does not go to zero, the gap can usually not be expected to vanish either. However, we can still show decrease in the gap up to a residual stemming from the variance. In particular, for a constant stepsize $\alpha_k = \alpha$ we have

$$\mathbb{E}[G(\bar{w}_K)] \leq \frac{D^2}{2\alpha K} + 9\sigma^2\alpha. \tag{5.17}$$

Additionally, if the number of iterations K is fixed beforehand, a conclusion similar to (5.16) can be obtained by choosing $\alpha = 1/\sqrt{K}$ in (5.17).

5.3.4 Proofs

We introduce the notation connected to the strong formulation of the VI (5.14) associated to the monotone inclusion (5.9), given by

$$g(w,z) := \langle F(w), w - z\rangle + r(w) - r(z),$$

for $g : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$. Next we will establish the fact that this function can be used to bound the (restricted) unifying gap function, which we remind, is defined as

$$G_B(w) = \begin{cases} \sup_{(x,y)\in B} \Psi(u,y) - \Psi(x,v) & \text{if } F \text{ is (5.2)} \\ \sup_{z\in B} \langle F(z), w - z\rangle + r(w) - r(z) & \text{otherwise,} \end{cases}$$

where in the first case $(u,v) \in \mathbb{R}^d \times \mathbb{R}^n$ is identified with $w \in \mathbb{R}^m$. In particular the dimensions fulfill $d + n = m$, and $r(w)$ is given by $f(u) + h(v)$.

Lemma 5.3.14. *It holds that for all $K \geq 1$*

$$\sup_{z\in B} \left\{ \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \sum_{k=0}^{K-1} \alpha_k g(w_k, z) \right\} \geq G_B(\bar{w}_K).$$

Proof. First we will prove the case if F is derived from a saddle point problem. Note that from the convex-concave structure of Φ we get that

$$\Phi(u,y) \leq \Phi(u,v) + \langle \nabla_y \Phi(u,v), y - v\rangle$$

and

$$\Phi(u,v) + \langle \nabla_x \Phi(u,v), x - u\rangle \leq \Phi(x,v).$$

By summing the two up we obtain

$$\Phi(u,y) - \Phi(x,v) \leq \left\langle \begin{matrix} -\nabla_x \Phi(u,v), \\ \nabla_y \Phi(u,v), \end{matrix} \begin{matrix} x - u \\ y - v \end{matrix} \right\rangle.$$

We can reformulate the above inequality in terms of g to see that for $z = (x,y) \in \mathbb{R}^d \times \mathbb{R}^n$

$$\langle F(w), w - z\rangle \geq \Phi(u,y) - \Phi(x,v).$$

The statement of the first case is obtained by adding $r(w) - r(z)$ on both sides and using the fact that Ψ is convex-concave.

If F is a general monotone operator, then we use its monotonicity to deduce that

$$\langle F(w), w - z\rangle \geq \langle F(z), w - z\rangle.$$

The desired result follows from using the linearity of the inner product. □

Notation. We denote the error of the stochastic estimator via

$$Z_k := F(\diamondsuit_k; \triangle_k) - F(\diamondsuit_k) \quad \text{and} \quad W_k := F(w_k; \xi_k) - F(w_k). \tag{5.18}$$

Furthermore, we will denote via $\mathbb{E}[\,\cdot \,|\, U]$, the conditional expectation with respect to the random variable U.

A unified decrease result

We will start with a unifying proposition which covers the common parts of all convergence proofs.

Proposition 5.3.15. *For a $\gamma > 0$ we have for all $k \geq 0$ and $z \in \mathbb{R}^m$*

$$\begin{aligned} &\alpha_k \mathbb{E}[g(w_k, z)] + \frac{1}{2}\mathbb{E}\|z_{k+1} - z\|^2 \\ &\quad \leq \frac{1}{2}\mathbb{E}\|z_k - z\|^2 - \frac{1}{2}\mathbb{E}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \mathbb{E}\|\diamondsuit_k - w_k\|^2 + 2(1+\gamma^{-1})\alpha_k^2\sigma^2. \end{aligned} \tag{5.19}$$

Proof. Let $k \geq 0$ and $z \in \mathbb{R}^m$ be arbitrary. Using the decomposition (5.18) it follows that

$$\langle \alpha_k F(w_k; \xi_k), w_k - z\rangle = \alpha_k \langle W_k, w_k - z\rangle + \alpha_k \langle F(w_k), w_k - z\rangle. \tag{5.20}$$

Since, from the proximal operator $w_k + \alpha_k \partial r(w_k) = z_k - \alpha_k F(\diamondsuit_k; \triangle_k)$ we deduce that

$$\langle z - w_k, w_k - z_k + \alpha_k F(\diamondsuit_k; \triangle_k)\rangle \geq \alpha_k (r(w_k) - r(z)). \tag{5.21}$$

Adding (5.20) and (5.21) gives that

$$\langle \alpha_k (F(w_k; \xi_k) - F(\diamondsuit_k; \triangle_k)) + z_k - w_k, w_k - z\rangle \geq \alpha_k \langle W_k, w_k - z\rangle + \alpha_k g(w_k, z),$$

which, using the definition of z_{k+1}, is equivalent to

$$\langle z - w_k, z_{k+1} - z_k\rangle \geq \alpha_k \langle W_k, w_k - z\rangle + \alpha_k g(w_k, z). \tag{5.22}$$

We estimate the inner product on the left side of the inequality by inserting and subtracting z_k and using the three point identity twice to deduce

$$\begin{aligned} \langle z - w_k, z_{k+1} - z_k\rangle &= \langle z - z_k + z_k - w_k, z_{k+1} - z_k\rangle \\ &= \frac{1}{2}\left(\|z - z_k\|^2 - \|z_{k+1} - z\|^2 + \|z_{k+1} - w_k\|^2 - \|z_k - w_k\|^2\right). \end{aligned} \tag{5.23}$$

The first two summands are fine as they will telescope, so we are left with estimating $\|z_{k+1} - w_k\|^2$. By the definition of z_{k+1} we have that

$$\begin{aligned} \|z_{k+1} - w_k\|^2 &= \alpha_k^2 \|F(\diamondsuit_k; \triangle_k) - F(w_k; \xi_k)\|^2 \\ &= \alpha_k^2 \|F(\diamondsuit_k) - F(w_k) + Z_k - W_k\|^2 \\ &\leq (1+\gamma)\alpha_k^2 \|F(\diamondsuit_k) - F(w_k)\|^2 + (1+\gamma^{-1})\alpha_k^2 \|Z_k - W_k\|^2 \\ &\leq (1+\gamma)\alpha_k^2 L^2 \|\diamondsuit_k - w_k\|^2 + 2(1+\gamma^{-1})\alpha_k^2 (\|Z_k\|^2 + \|W_k\|^2), \end{aligned} \tag{5.24}$$

where we inserted and subtracted $F(\diamondsuit_k)$ and $F(w_k)$ and applied Young's inequality to deduce. Adding (5.24), (5.23) and (5.22) we deduce that

$$\begin{aligned}\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \|\diamondsuit_k - w_k\|^2 \\ + \alpha_k \langle W_k, z - w_k\rangle + (1+\gamma^{-1})\alpha_k^2(\|W_k\|^2 + \|Z_k\|^2).\end{aligned}$$

Taking the expectation $\mathbb{E}[\cdot]$ and using the bounded variance assumption of the estimators yields

$$\begin{aligned}&\alpha_k \mathbb{E}[g(w_k, z)] + \frac{1}{2}\mathbb{E}\|z_{k+1} - z\|^2 \\ &\quad\leq \frac{1}{2}\mathbb{E}\|z_k - z\|^2 - \frac{1}{2}\mathbb{E}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \mathbb{E}\|\diamondsuit_k - w_k\|^2 + 2(1+\gamma^{-1})\alpha_k^2\sigma^2,\end{aligned}$$

where we used that

$$\mathbb{E}[\langle W_k, z - w_k\rangle] = \mathbb{E}\Big[\mathbb{E}[\langle W_k, z - w_k\rangle \,|\, w_k]\Big] = \mathbb{E}\Big[\langle \mathbb{E}[W_k \,|\, w_k], z - w_k\rangle\Big] = \mathbb{E}[0] = 0,$$

since

$$\mathbb{E}[W_k \,|\, w_k] = \mathbb{E}[F(w_k; \xi_k) - F(w_k) \,|\, w_k] \overset{(*)}{=} F(w_k) - F(w_k) = 0.$$

Here, $(*)$ holds because of the independence and unbiasedness, see Assumption 5.3.12 and 5.3.10, respectively. □

Forward-Backward-Forward

Proof for deterministic FBF, Theorem 5.3.9 (i). We start off by plugging $\diamondsuit_k = z_k$ into (5.19). Since $\sigma = 0$ we can discard the expectations and use $\gamma \to 0$ to deduce that for all $k \geq 0$

$$\alpha_k g(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}(1 - \alpha_k^2 L^2)\|z_k - w_k\|^2.$$

From this it is clear that the stepsize is constrained by $\alpha \leq 1/L$ as stated in the theorem. By summing up from $k = 0$ to $K-1$ and dividing by $\sum_{k=0}^{K-1}\alpha_k$ we obtain

$$\frac{1}{\sum_{k=0}^{K-1}\alpha_k}\sum_{k=0}^{K-1}\alpha_k g(w_k, z) \leq \frac{\|z_0 - z\|^2}{2\sum_{k=0}^{K-1}\alpha_k}.$$

The claimed statement is then derived by taking the supremum in z over B and applying Lemma 5.3.14. □

Proof for stochastic FBF, Theorem 5.3.13 (i). Plugging $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$ into (5.19) gives for all $k \geq 0$

$$\begin{aligned}&\alpha_k \mathbb{E}[g(w_k, z)] + \frac{1}{2}\mathbb{E}\|z_{k+1} - z\|^2 \\ &\quad\leq \frac{1}{2}\mathbb{E}\|z_k - z\|^2 - \frac{1}{2}(1 - (1+\gamma)\alpha_k^2 L^2)\mathbb{E}\|z_k - w_k\|^2 + 2(1+\gamma^{-1})\alpha_k^2\sigma^2.\end{aligned}$$

By choosing γ such that $\alpha = (\sqrt{1+\gamma}L)^{-1}$ we deduce that $1+\gamma^{-1} = 1/(1-\alpha^2L^2)$. Next, we sum up and divide by $\sum_{k=0}^{K-1}\alpha_k$ to obtain

$$\mathbb{E}\left[\frac{1}{\sum_{k=0}^{K-1}\alpha_k}\sum_{k=0}^{K-1}\alpha_k g(w_k,z)\right] \le \frac{\mathbb{E}\|z_0-z\|^2+4(1-\alpha^2L^2)^{-1}\sigma^2\sum_{k=0}^{K-1}\alpha_k^2}{2\sum_{k=0}^{K-1}\alpha_k}.$$

The final statement follows by taking the supremum in z over B and applying Lemma 5.3.14. □

Forward-backward-forward-past

Proof for deterministic FBFp, Theorem 5.3.9 (ii). We start off by plugging $\diamondsuit_k = z_k$ into (5.19). Since $\sigma = 0$ we can ignore the expectations and use $\gamma \to 0$ to conclude that for all $k \ge 0$

$$\alpha_k g(w_k,z) + \frac{1}{2}\|z_{k+1}-z\|^2 \le \frac{1}{2}\|z_k-z\|^2 - \frac{1}{2}\|z_k-w_k\|^2 + \frac{1}{2}\alpha_k^2L^2\|w_{k-1}-w_k\|^2. \tag{5.25}$$

Now we need to bound the term $\|w_{k-1}-w_k\|^2$ by $\|z_k-w_k\|^2$. Since

$$2\|z_k-w_k\|^2+2\|z_k-w_{k-1}\|^2 \ge \|w_k-w_{k-1}\|^2 \tag{5.26}$$

we have for all $k \ge 1$

$$\begin{aligned}\|z_k-w_k\|^2 &\ge -\|z_k-w_{k-1}\|^2+\frac{1}{2}\|w_{k-1}-w_k\|^2\\ &\ge -\alpha_{k-1}^2L^2\|w_{k-1}-w_{k-2}\|^2+\frac{1}{2}\|w_{k-1}-w_k\|^2\end{aligned} \tag{5.27}$$

whereas for $k = 0$, since $w_{-1} = z_0$, we have that

$$\|z_0-w_0\|^2 = \|w_{-1}-w_0\|^2. \tag{5.28}$$

Plugging (5.28) into (5.25) for $k = 0$ we get that

$$\alpha_0 g(w_0,z) + \frac{1}{2}\|z_1-z\|^2+\frac{1}{2}(1-\alpha_0^2L^2)\|w_0-w_{-1}\|^2 \le \frac{1}{2}\|z_0-z\|^2. \tag{5.29}$$

Plugging (5.27) into (5.25) we get that for all $k \ge 1$

$$\begin{aligned}&\alpha_k g(w_k,z) + \frac{1}{2}\|z_{k+1}-z\|^2+\frac{1}{2}\left(\frac{1}{2}-\alpha_k^2L^2\right)\|w_k-w_{k-1}\|^2\\ &\le \frac{1}{2}\|z_k-z\|^2+\frac{1}{2}\alpha_{k-1}^2L^2\|w_{k-1}-w_{k-2}\|^2.\end{aligned} \tag{5.30}$$

In order to be able to telescope we need to ensure that for all $k \ge 0$

$$\left(\frac{1}{2}-\alpha_k^2L^2\right) \ge \alpha_k^2L^2.$$

This is equivalent to the condition $\alpha_k \leq {}^{1}\!/_{2L}$ which was required in the statement of the theorem. Now we sum up (5.30) from $k = 1$ to $K - 1$ which yields

$$\begin{aligned}\sum_{k=1}^{K-1} \alpha_k g(w_k, z) + \frac{1}{2}\|z_K - z\|^2 + \frac{1}{2}\left(\frac{1}{2} - \alpha_{K-1}^2 L^2\right)\|w_{K-1} - w_{K-2}\|^2 \\ \leq \frac{1}{2}\|z_1 - z\|^2 + \frac{1}{2}\alpha_0^2 L^2 \|w_0 - w_{-1}\|^2.\end{aligned} \tag{5.31}$$

Adding (5.31) and (5.29) and dividing by $\sum_{k=0}^{K-1} \alpha_k$ to deduce

$$\frac{1}{\sum_{k=0}^{K-1} \alpha_k} \sum_{k=0}^{K-1} \alpha_k g(w_k, z) \leq \frac{\|z_0 - z\|^2}{2\sum_{k=0}^{K-1} \alpha_k},$$

where we used that $1 - \alpha_0^2 L^2 \geq \alpha_0^2 L^2$ to get rid of $\|w_0 - w_{-1}\|^2$. The final statement follows by taking the supremum in z over B and applying Lemma 5.3.14. □

Proof for stochastic FBFp, Theorem 5.3.13 (ii). By using $\diamondsuit_k = w_{k-1}$ we deduce from (5.19) for all $k \geq 0$ that

$$\begin{aligned}&\alpha_k \mathbb{E}[g(w_k, z)] + \frac{1}{2}\mathbb{E}\|z_{k+1} - z\|^2 \\ &\quad \leq \frac{1}{2}\mathbb{E}\|z_k - z\|^2 - \frac{1}{2}\mathbb{E}\|z_k - w_k\|^2 + \frac{1}{2}(1+\gamma)\alpha_k^2 L^2 \mathbb{E}\|w_{k-1} - w_k\|^2 + 2(1+\gamma^{-1})\alpha_k^2 \sigma^2.\end{aligned} \tag{5.32}$$

Let from now on $k \geq 1$ as we will treat the case $k = 0$ separately. Using (5.26) we deduce that

$$\begin{aligned}\|z_k - w_k\|^2 &\geq -\|z_k - w_{k-1}\|^2 + \frac{1}{2}\|w_{k-1} - w_k\|^2 \\ &\geq -\alpha_{k-1}^2 \|F(w_{k-1}; \xi_{k-1}) - F(w_{k-2}; \xi_{k-2})\|^2 + \frac{1}{2}\|w_{k-1} - w_k\|^2.\end{aligned} \tag{5.33}$$

Now we bound the difference of the two estimators by inserting $\pm F(w_{k-1})$, $\pm F(w_{k-2})$ and applying the inequality $\|a + b + c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$ which yields

$$\begin{aligned}&\|F(w_{k-1}; \xi_{k-1}) - F(w_{k-2}; \xi_{k-2})\|^2 \\ &\quad \leq 3\|W_{k-1}\|^2 + 3\|W_{k-2}\|^2 + 3\|F(w_{k-2}) - F(w_{k-1})\|^2.\end{aligned}$$

We conclude that

$$\mathbb{E}\big[\|F(w_{k-1}; \xi_{k-1}) - F(w_{k-2}; \xi_{k-2})\|^2\big] \leq 6\sigma^2 + 3L^2 \mathbb{E}\|w_{k-1} - w_{k-2}\|^2. \tag{5.34}$$

Using (5.34) in (5.33) we deduce that

$$\mathbb{E}\|z_k - w_k\|^2 \geq -\alpha_{k-1}^2 (6\sigma^2 + 3L^2 \mathbb{E}\|w_{k-1} - w_{k-2}\|^2) + \frac{1}{2}\mathbb{E}\|w_{k-1} - w_k\|^2, \tag{5.35}$$

whereas for $k = 0$ we have (5.28). Now we plug (5.35) into (5.32) to conclude that

$$\begin{aligned}\alpha_k \mathbb{E}[g(w_k, z)] + \frac{1}{2}\mathbb{E}\|z_{k+1} - z\|^2 + \frac{1}{2}\left(\frac{1}{2} - (1+\gamma)\alpha_k^2 L^2\right)\mathbb{E}\|w_k - w_{k-1}\|^2 \\ \leq \frac{1}{2}\mathbb{E}\|z_k - z\|^2 + \frac{1}{2}3\alpha_{k-1}^2 L^2 \mathbb{E}\|w_{k-1} - w_{k-2}\|^2 + (2(1+\gamma^{-1})\alpha_k^2 + 3\alpha_{k-1}^2)\sigma^2.\end{aligned} \tag{5.36}$$

From this we conclude that in order to be able to telescope we need to enforce

$$\left(\frac{1}{2} - (1+\gamma)\alpha_k^2 L^2\right) \geq 3\alpha_k^2 L^2,$$

which is equivalent to

$$\frac{1}{2(4+\gamma)} \geq \alpha_k^2 L^2.$$

Since $\alpha_k \leq \alpha$, we can ensure this by choosing γ such that

$$\frac{1}{2(4+\gamma)} = \alpha^2 L^2. \tag{5.37}$$

With (5.37) in place we sum (5.36) from $k = 1$ to $K - 1$ to deduce that

$$\begin{aligned}\sum_{k=1}^{K-1} \alpha_k \mathbb{E}[g(w_k, z)] + \frac{1}{2}\mathbb{E}\|z_K - z\|^2 + \frac{1}{2}\left(\frac{1}{2} - (1+\gamma)\alpha_{K-1}^2 L^2\right)\mathbb{E}\|w_{K-1} - w_{K-2}\|^2 \\ \leq \frac{1}{2}\mathbb{E}\|z_1 - z\|^2 + \frac{1}{2}3\alpha_0^2 L^2 \|w_0 - w_{-1}\|^2 + (5 + 2\gamma^{-1})\sigma^2 \sum_{k=1}^{K-1} \alpha_k^2 + 3\sigma^2\alpha_0^2,\end{aligned} \tag{5.38}$$

whereas for $k = 0$ we have

$$\alpha_0 \mathbb{E}[g(w_0, z)] + \frac{1}{2}\mathbb{E}\|z_1 - z\|^2 + \frac{1}{2}(1 - (1+\gamma)\alpha_0^2 L^2)\mathbb{E}\|w_0 - w_{-1}\|^2 \leq \frac{1}{2}\|z_0 - z\|^2 + 2(1+\gamma^{-1})\alpha_0^2\sigma^2. \tag{5.39}$$

Combining (5.38) and (5.39) and using the fact that $3\alpha_0^2 L^2 \leq 1 - (1+\gamma)\alpha_0^2 L^2$ from (5.37) to discard the $\|w_0 - w_{-1}\|^2$ term, yields

$$\sum_{k=0}^{K-1} \alpha_k \mathbb{E}[g(w_k, z)] \leq \frac{1}{2}\|z_0 - z\|^2 + (5 + 2\gamma^{-1})\sigma^2 \sum_{k=0}^{K-1} \alpha_k^2. \tag{5.40}$$

Through (5.37), we can estimate

$$\frac{1}{\gamma} = \frac{2\alpha^2 L^2}{1 - 8\alpha^2 L^2}. \tag{5.41}$$

Plugging (5.41) into (5.40), dividing by $\sum_{k=0}^{K-1} \alpha_k$ taking the supremum in z over B and applying Lemma 5.3.14, deduces the final statement. □

5.4 Experiments

Due to the theoretical nature of this work, the aim of this section is rather to validate the results on standard examples and not to strive to achieve new state-of-the-art results. Instead we simply aim to show how the use of methods with convergence guarantees, albeit only in the monotone setting, can yield better training performance.

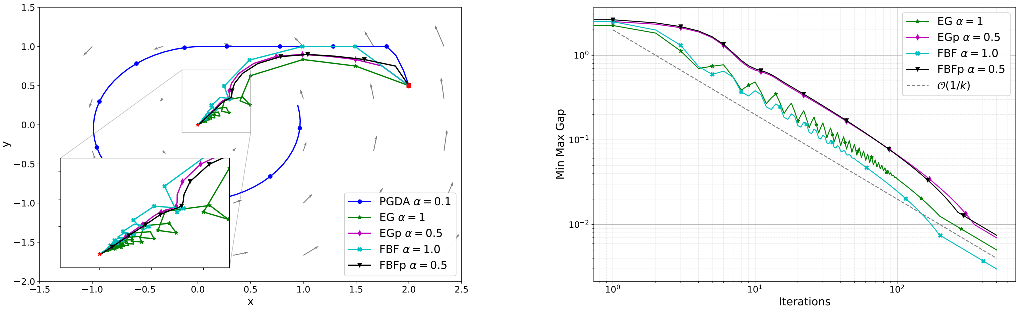
5.4.1 2D toy example

Following [40, 42, 68] we consider the canonical example $\min_x \max_y xy$, which illustrates the cycling behavior of (even bilinear) minimax problems, and augment this approach by adding a nonsmooth L1-regularizer for one player, resulting in

$$\min_{x\in\mathbb{R}} \max_{y\in[-1,1]} \kappa|x|+xy, \tag{5.42}$$

with $\kappa > 0$.

Figure 5.1 highlights the aforementioned issue of GDA (and its proximal extension PGDA) cycling around the solution. The other methods, for which we display the averaged iterates, however do converge to a solution and show a decrease in the restricted gap according to theory. Even though the proximal steps provide improvement towards the solution $(0, 0)$ and FBF only uses half the amount of evaluations compared to EG, it outperforms the competing algorithms.



(a) Trajectories converging to solution.

(b) Restricted gap function.

Figure 5.1: A comparison of the methods presented in Section 5.2.3 applied to problem (5.42) with $\kappa = 0.01$. *PGDA* denotes (alternating) gradient descent ascent with proximal steps. As mentioned in the introduction it fails to converge. *EGp* denotes the method presented in [40] as extrapolation from the past. For the restricted gap we use $B_1 = B_2 = [-1, 1]$.

5.4.2 WGAN trained on CIFAR10

In this section we apply the above proposed techniques from monotone inclusions to the training of Wasserstein GANs making use of the DCGAN architecture [85]. All models are trained on the CIFAR10 dataset [55] which consists of 60,000 images in 10 different classes (with 50,000 training images and 10,000 test images) using an NVIDIA RTX 2080Ti GPU.

We choose to work with the original WGAN formulation including weight clipping, since it includes regularizers innately (the indicator of a box for the weights of the discriminator). Although more recent models like ones for example based on ResNet [47] or SAGAN [109] architectures provide better overall performance, they usually do not warrant the use of regularizers. We do this to highlight the difference between FBF and EG, as without projections or proximal steps they are equivalent and their relevance including state-of-the-art architectures has already been shown [29, 40].

In addition we propose a modification of the WGAN formulation which replaces the box constraint on the discriminator's weights with an L1-regularization, under the name of *WGAN-L1*. This results in a *soft-thresholding* operation instead of the "harsh" clipping.

	Inception Score (IS)		Fréchet Inception Distance (FID)	
Method	clip	prox	clip	prox
AltAdam1	4.12±0.06	4.43±0.03	56.44±0.62	50.86±2.17
Extra Adam	4.07±0.05	4.67±0.11	56.67±0.61	47.24±1.21
FBF Adam	**4.54±0.04**	**4.68±0.16**	**45.85±0.35**	**46.60±0.76**
Optimistic Adam	4.35±0.06	4.63±0.13	50.41±0.46	47.98±1.49

Table 5.1: The best Inception Score (IS) and Fréchet Inception Distance (FID), averaged over 5 runs. The column denoted by *clip* refers the standard formulation WGANs where the weights of the discriminator are clipped after every gradient step to enforce the box constraint, whereas *prox* refers alternative implementation using the 1-norm of the weights for regularization. The latter provides improvement throughout all considered methods. For both formulations, the FBF method (with Adam update) yields the best results (higher IS and lower FID).

Given the ubiquity and dominance of Adam [52] as an optimizer for many deep learning related training tasks, instead of using vanilla SGD we opt for Adam updates. This results in a method we call *FBF Adam*. Analogous approaches have been applied in [40] and [33] resulting in *Extra Adam* and *Optimistic Adam*, respectively. We compare the aforementioned methods with the status-quo in GAN training, namely alternating one Adam step for each network: *AltAdam1*.

Our hyperparameter search was limited to the stepsizes when using the WGAN-L1 formulation, while all other parameters were kept the same as in [24, 40]. It seems noteworthy that in the case of soft-thresholding bigger stepsizes performed better with the only exception of AltAdam1.

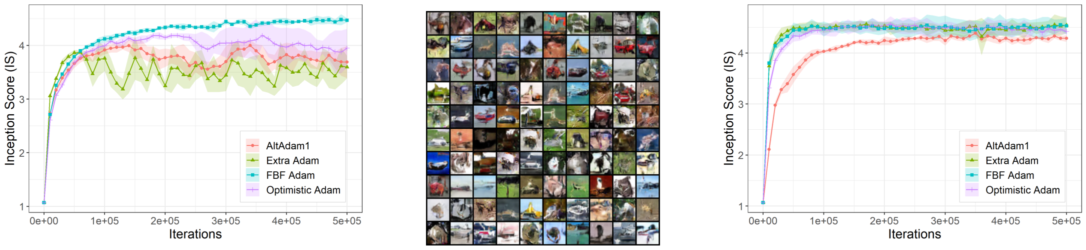


Figure 5.2: **Left:** Mean and standard deviation of the IS averaged over 5 runs on the WGAN objective with weight clipping. **Middle:** Samples from the DCGAN generator trained with the WGAN-L1 objective using the FBF method with Adam updates. **Right:** Mean and standard deviation of the IS averaged over 5 runs on the WGAN-L1 objective using the proximal operator; The WGAN-L1 objective improves the IS in comparison to weight clipping and stabilizes the behavior of all considered methods during the training procedure. The advantage of using FBF Adam is most pronounced in the case of weight clipping.

The two evaluation metrics used are the Inception Score (IS) [95] and the Fréchet inception distance (FID) [48], both computed on 50,000 samples. In the case of the IS we use the updated and corrected implementation from [5]. All results are averaged over 5 runs for each method.

Table 5.1 reports the best IS and FID for each method. FBF Adam outperforms all considered competitors with respect to both evaluation metrics with the most significant difference for WGAN with weight clipping (“clip”). One can also see that WGAN-L1 using the proximal operator (“prox”) improves the performance of all considered methods, decreasing the absolute and relative differences. Note that the results with WGAN-L1 are comparable for the three methods with underlying convergence guarantees in the convex-concave case. Figure 5.2 shows the training progress regarding IS for each method and both problem formulations. The graphs suggest that making use of WGAN-L1 objective has a stabilizing effect during training leading to a smoother and more consistent learning curve — a property that only FBF Adam seems to exhibit for weight clipping.

6 Weakly convex-concave minimax problems

Minimax problems of the form $\min_x \max_y \Psi(x, y)$ have attracted increased interest largely due to advances in machine learning, in particular generative adversarial networks. These are typically trained using variants of stochastic gradient descent for the two players. Although convex-concave problems are well understood with many efficient solution methods to choose from, theoretical guarantees outside of this setting are sometimes lacking even for the simplest algorithms. In particular, this is the case for alternating gradient descent ascent, where the two agents take turns updating their strategies. To partially close this gap in the literature we prove a novel global convergence rate for the stochastic version of this method for finding a critical point of $g(\cdot) := \max_y \Psi(\cdot, y)$ in a setting which is not convex-concave.

6.1 Introduction

We investigate the *alternating* proximal gradient descent ascent (GDA) method for weakly convex-(strongly) concave saddle point problems, given by

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^n} \Big\{\Psi(x, y) := f(x) + \Phi(x, y) - h(y)\Big\} \tag{6.1}$$

for a weakly convex-concave coupling function $\Phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ and proper, convex and lower semicontinuous regularizers $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, see Assumption 6.3.1, 6.3.3 and 6.4.1 for details.

Nonconvex-concave saddle point problems have received a great deal of attention in the recently due to their application in adversarial learning [98], learning with nondecomposable losses [39, 107], and learning with uncertain data [31]. Additionally, albeit typically resulting in intricate nonconvex-nonconcave objectives, the large interest in *generative adversarial networks (GANs)* [2, 43] has led to the studying of saddle point problems under different simplifying assumptions [4, 12, 33, 40, 60].

In the nonconvex-concave setting *inner loop* methods have received much of the attention [53, 60, 70, 80, 99] with them obtaining the best complexity results in this class, see Table 6.1. Despite superior theoretical performance these methods have not been as popular in practice, especially in the training of GANs where *single loop* methods are still state-of-the-art [4, 33, 40, 41, 43, 49, 62]. The simplest approach is given by *simultaneous* GDA, which, for a smooth coupling function Φ and stepsizes $\eta_x, \eta_y > 0$, reads as:

$$\text{(simultaneous)} \quad \left\lfloor \begin{array}{l} x^+ = x - \eta_x \nabla_x \Phi(x, y) \\ y^+ = y + \eta_y \nabla_y \Phi(x, y). \end{array} \right.$$

After the first step of this method, however, more information is already available, which can be used in the update of the second variable, resulting in

$$\text{(alternating)} \quad \left\lfloor \begin{array}{l} x^+ = x - \eta_x \nabla_x \Phi(x,y) \\ y^+ = y + \eta_y \nabla_y \Phi(x^+,y). \end{array} \right.$$

It has been widely known that the alternating version of GDA has many favorable convergence properties of the simultaneous one [4, 41, 106]. We are naturally interested in — and will give an affirmative answer to the question:

Does stochastic alternating GDA have nonasymptotic convergence guarantees for nonconvex minimax problems?

This might seem surprising as it has been sufficiently demonstrated [12, 40, 42, 68] that either version of GDA fails to converge if equal stepsizes are used. We therefore want to point out the importance of the *two-time-scale* approach which was also emphasized in [48, 60].

Optimality. For convex-concave minimax problems, the notion of solution is simple. We aim to find a so-called saddle point $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^n$ satisfying for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^n$

$$\Psi(x^*, y) \leq \Psi(x^*, y^*) \leq \Psi(x, y^*). \tag{6.2}$$

For convex-concave problems this notion is equivalent to the first order optimality condition

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \nabla_x \Phi(x^*, y^*) \\ -\nabla_y \Phi(x^*, y^*) \end{pmatrix} + \begin{pmatrix} \partial f(x^*) \\ \partial h(y^*) \end{pmatrix}. \tag{6.3}$$

Similarly to the nonconvex single objective optimization where one cannot expect to find global minima, if the minimax problem is not convex-concave the notion of saddle point is too strong. So one natural approach is to focus on conditions such as (6.3), as done in [64, 78, 106]. However, treating the two components in such a symmetric fashion might not seem fitting since in contrast to the convex-concave problem $\min_x \max_y \neq \max_y \min_x$. Instead we will focus, in the spirit of [60, 86, 99], on the stationarity of what we will refer to as the *max function* given by

$$\varphi(x) := \max_{y \in \mathbb{R}^n} \Phi(x, y) - h(y), \quad \text{where } \varphi : \mathbb{R}^d \to \mathbb{R}. \tag{6.4}$$

This makes sense from the point of view of many practical applications. Problems arising from adversarial learning can be formulated as minimax, but typically only x, which corresponds to the classifier is relevant as y is adversarial noise. Similarly, for GANs, one is typically only interested in the generator and not the discriminator. See Table 6.1 for a comparison of other methods using the same notion of optimality. Note that it is possible to move from one notion of optimality to the other [60], but as both directions are typically associated with additional computational effort a comparison is not trivial and out of scope of this work.

Table 6.1: The gradient complexity of algorithms for nonconvex-(strongly) concave minimax problems and their convergence rates for (near) stationarity of the max function. ϵ is the tolerance and $\kappa > 0$ is the condition number.

	Nonconvex-Strongly Concave		**Nonconvex-Concave**		single
	Deterministic	Stochastic	Deterministic	Stochastic	loop
[86]	$\tilde{\mathcal{O}}(\kappa^2\epsilon^{-2})$	$\tilde{\mathcal{O}}(\kappa^3\epsilon^{-4})$	$\tilde{\mathcal{O}}(\epsilon^{-6})$	$\tilde{\mathcal{O}}(\epsilon^{-6})$	✗
[99, 110]	–	–	$\tilde{\mathcal{O}}(\epsilon^{-3})$	–	✗
[61, 80]	$\tilde{\mathcal{O}}(\sqrt{\kappa}\epsilon^{-2})$	–	$\tilde{\mathcal{O}}(\epsilon^{-3})$	–	✗
[60]	$\mathcal{O}(\kappa^2\epsilon^{-2})$	$\mathcal{O}(\kappa^3\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-8})$	✓
this work	$\mathcal{O}(\kappa^2\epsilon^{-2})$	$\mathcal{O}(\kappa^3\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{O}(\epsilon^{-8})$	✓

Contributions. We prove novel convergence rates for *alternating* gradient descent ascent for nonconvex-(strongly) concave minimax problems in a deterministic and stochastic setting. For deterministic problems, [106] has proved convergence rates for alternating GDA in terms of the criticality of Φ while we use the max function φ, see (6.4), instead. Our results are also more general than e.g. [60, 61, 110] in the sense that they require Φ to be smooth in the first component wheres we only require weak convexity, similar to [86]. Furthermore, we allow for our method to include possibly nonsmooth regularizers, similar to [86, 110], which captures and extends the common constraint setting.

6.1.1 Related literature

For the purpose of this paper we separate the nonasymptotic study of minimax problems into the following domains.

Convex-concave. For convex-concave problems historically the *extra-gradient* and the *forward-backward-forward* method have been known to converge. For the former even a rate of $\mathcal{O}(\epsilon^{-1})$ has been proven in [71] under the name of *mirror-prox*. Both of these methods suffer from the drawback of requiring two gradient evaluations per iteration. This has led to the development fo methods such as *optimistic GDA* [33, 34] or [12, 40, 46, 66] which use past gradients to reduce the need of gradient evaluations to one per iteration. In all of these cases, however, convergence guarantees typically do not go beyond the convex-concave setting. Nevertheless, these methods have been employed successfully in the GAN setting [12, 33, 40].

Nonconvex-concave with inner loops. Approximating the max function by running multiple iterations of a solver on the second component or convexifying the problem by adding a quadratic term and then solving the convex-concave problem constitute natural approaches [61, 78, 86, 99, 110]. Such methods achieve the best known rates [61, 80, 99, 110] in this class. However, they are usually quite involved and have for the most part not been used in deep learning applications.

Nonconvex-concave with single loop. While these methods have received some attention in the training of GANs [12,33,40] most of the theoretical statement are for convex-concave problems. In the nonconvex setting only two methods have been studied. Previous research, see [60,64], has focused on the *simultaneous* version of the gradient descent ascent algorithm where both components are updated at the same time. The only other work which focuses on *alternating* GDA is [106]. Their results are in terms of stationarity of Φ and they do not treat the stochastic case. Note that our work is most similar to [60] where the same notion of optimality is used and similar rates to our are obtained for *simultaneous* GDA.

Others. Clearly the above categories do not cover the entire field. However, other settings have not received as much attention. Only [106] treats (strongly) convex-nonconcave problems and proves convergence rates similar to the nonconvex-(strongly) concave setting. In [102] a special stochastic nonconvex-linear problem with regularizers is solved via a variance reduced single loop method with a significantly improved rate over the general nonconvex-concave problem.

The most general setting out of all the aforementioned ones is discussed in [59, 62, 67], namely the weakly convex-weakly concave setting. They use however, a weaker notion of optimality related to the Minty variational inequality formulation. We also only mentioned (sub)gradient methods, but the restrictive assumption that the proximal operator of a component can be evaluated has been considered as well [53].

6.2 Preliminaries

As mentioned in the earlier we will consider optimality in terms of the max function for any $x \in \mathbb{R}^d$ given by $\varphi(x) := \max_{y \in \mathbb{R}^n} \psi(x, y)$, for $\psi(x, y) := \Phi(x, y) - h(y)$ as mapping from $\mathbb{R}^d \times \mathbb{R}^n$ to $\mathbb{R} \cup \{-\infty\}$. Similarly, we also need the regularized max function

$$g := \varphi + f, \quad \text{where } g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}.$$

In the remainder of the section we will focus on the necessary preliminaries connected to the weak convexity of the max function in the nonconvex-concave setting, see Section 6.3.

6.2.1 About the stochastic setting

We discuss the stochastic version of problem (6.1) where the coupling function Φ is actually given as an expectation, i.e.

$$\Phi(x, y) = \mathbb{E}_{\xi \sim \mathcal{D}} [\Phi(x, y; \xi)] \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^n$$

and we can only access independent samples of the gradient $\nabla_x \Phi(x, y; \xi)$ (or subgradient) and $\nabla_y \Phi(x, y; \zeta)$, where ξ and ζ are drawn from the (in general unknown) distribution $\mathcal{D}$.

We require the following standard assumption with respect to these stochastic gradient estimators.

Assumption 6.2.1 (unbiased)**.** *The stochastic estimator of the gradient is unbiased, i.e. for all* $(x, y) \in \mathbb{R}^d \times \mathbb{R}^n$

$$\mathbb{E}[\nabla\Phi(x, y; \xi)] = \nabla\Phi(x, y),$$

or in the case of subgradients

$$\mathbb{E}\Big[g^\xi\Big] \in \partial[\Phi(\cdot, y)](x), \quad \textit{where } g^\xi \in \partial[\Phi(\cdot, y; \xi)](x).$$

Assumption 6.2.2 (bounded variance)**.** *The variance of the estimator is uniformly bounded, i.e. for all* $(x, y) \in \mathbb{R}^d \times \mathbb{R}^n$

$$\mathbb{E}\big[\|\nabla_x\Phi(x, y; \xi) - \nabla_x\Phi(x, y)\|^2\big] \leq \sigma^2 \quad \textit{and} \quad \mathbb{E}\big[\|\nabla_y\Phi(x, y; \xi) - \nabla_y\Phi(x, y)\|^2\big] \leq \sigma^2, \tag{6.5}$$

for a variance $\sigma^2 \geq 0$*. In the setting of Section 6.3 where* Φ *is not necessarily smooth in the first component, we make the analogous assumption for subgradients, i.e.*

$$\mathbb{E}\left[\left\|g^\xi - \mathbb{E}\Big[g^\xi\Big]\right\|^2\right] \leq \sigma^2 \tag{6.6}$$

for a stochastic subgradient $g^\xi \in \partial[\Phi(\cdot, y; \xi)](x)$*.*

6.2.2 The algorithm

Since we cover different settings such as smooth or not, deterministic and stochastic we try to formulate a unifying scheme.

Algorithm 6.2.3 (proximal alternating GDA)**.** Let $(x_0, y_0) \in \mathbb{R}^d \times \mathbb{R}^n$ and stepsizes $\eta_x, \eta_y > 0$. Consider the following iterative scheme

$$(\forall k \geq 0) \qquad \left\lfloor \begin{array}{l} x_{k+1} = \operatorname{prox}_{\eta_x f}\left(x_k - \eta_x G_x(x_k, y_k)\right) \\ y_{k+1} = \operatorname{prox}_{\eta_y h}\left(y_k + \eta_y G_y(x_{k+1}, y_k)\right), \end{array} \right.$$

where G_x and G_y will be replaced by the appropriate (sub)gradient and its estimator in the deterministic and stochastic setting, respectively.

6.3 Nonconvex-concave objective

In this section we treat the case where the objective function is weakly convex in x, but not necessarely smooth, and concave and smooth in y. This will result in a weakly convex max function whose Moreau envelope we will study for criticality.

6.3.1 Assumptions

While the first assumption concerns general setting of this section, i.e. weakly convex-concave, the latter assumptions are more of a technical nature.

Assumption 6.3.1. *The coupling function Φ is*

(i) concave and $L_{\nabla\Phi}$-smooth in the second component uniformly in the first one, i.e.

$$\|\nabla_y \Phi(x,y) - \nabla_y \Phi(x,y')\| \leq L_{\nabla\Phi}\|y-y'\| \quad \forall x \in \mathbb{R}^d \, \forall y, y' \in \mathbb{R}^n.$$

(ii) ρ-weakly convex in the first component uniformly in the second one, i.e.

$$\Phi(\cdot, y) + \frac{\rho}{2}\|\cdot\|^2 \quad \textit{is convex for all } y \in \mathbb{R}^n.$$

Assumption 6.3.1 is fulfilled if e.g. Φ is $L_{\nabla\Phi}$-smooth jointly in both components, i.e.

$$\|\nabla\Phi(x,y) - \nabla\Phi(x',y')\| \leq L_{\nabla\Phi}\|(x,y)-(x',y')\| \quad \forall x, x' \in \mathbb{R}^d \ \forall y, y' \in \mathbb{R}^n,$$

in which case (ii) holds with $\rho = L_{\nabla\Phi}$.

The next assumption is a classical technical assumption nonconvex optimization.

Assumption 6.3.2. *The function g is lower bounded, i.e.* $\inf_{x\in\mathbb{R}^d} g(x) > -\infty$.

In Section 6.3 we will actually need to bound the Moreau envelope g_λ, but these two conditions are in fact equivalent as for all $x \in \mathbb{R}^d$ and any $\lambda \in (0, \rho^{-1})$

$$g_\lambda(x) = \inf_{u\in\mathbb{R}^d} \left\{ g(u) + \frac{1}{2\lambda}\|x-u\|^2 \right\} \geq \inf_{u\in\mathbb{R}^d} g(u)$$

and conversely

$$g_\lambda(x) = \inf_{u\in\mathbb{R}^d} \left\{ g(u) + \frac{1}{2\lambda}\|x-u\|^2 \right\} \overset{u=x}{\leq} g(x).$$

We also want to point out that this assumption is weaker than the lower boundedness of Ψ, which is usually required if stationary points of the type (6.3) are used, see for example [64].

Assumption 6.3.3. *Φ is L-Lipschitz in the first component uniformly over* dom h *in the second one, i.e.*

$$\|\Phi(x,y) - \Phi(x',y)\| \leq L\|x-x'\| \quad \forall x, x' \in \mathbb{R}^d \ \forall y \in \operatorname{dom} h.$$

Assumption 6.3.4. *The regularizers f and h are proper, convex and lower semicontinuous.*

(i) Additionally, f is either L_f-Lipschitz continuous on its domain, which is assumed to be open, or the indicator of a nonempty, convex and closed set. Either of those assumptions guarantees for any $\gamma > 0$ the bound

$$\|\operatorname{prox}_{\gamma f}(x) - x\| \leq \gamma L_f \tag{6.7}$$

for all $x \in \operatorname{dom} f$ (in the case of the indicator the statement is trivially true).

(ii) Furthermore, h has bounded domain dom h *such that the diameter of* dom h *is bounded by C_h.*

6.3.2 Properties of the max function

Previous research, when concluding the weak convexity of the max function, has relied on the compactness of the domain over which to maximize. This is done so that the classical Danskin Theorem can be applied. This assumption is e.g. fulfilled in the context of *Wasserstein GANs* [2] with weight clipping, but not in other formulations such as [45]. We provide an extension of the classical Danskin Theorem, which only relies on the concavity and lower semicontinuity of the objective in the second component and the boundedness of $\operatorname{dom} h$, see Assumption 6.3.1 and 6.3.4. This implies that for every $x \in \mathbb{R}^d$ the set

$$Y(x) := \left\{ y^* \in \mathbb{R}^n : \varphi(x) = \Phi(x, y^*) - h(y^*) = \max_{y \in \mathbb{R}^n} \{\Phi(x, y) - h(y)\} \right\} \tag{6.8}$$

is nonempty. For brevity we will denote arbitrary elements of $Y(x_k)$ by y_k^* for all $k \geq 0$.

Proposition 6.3.5 (Subgradient characterization of the max function)**.** *Let Assumption 6.3.1 and 6.3.4 hold true. Then, the function φ, see* (6.4)*, fulfills for all $x \in \mathbb{R}^d$*

$$\partial[\Phi(\cdot, y^*)](x) \subseteq \partial\varphi(x) \quad \forall y^* \in Y(x).$$

In particular, φ is ρ-weakly convex.

Proof. From the ρ-weak convexity of $\Phi(\cdot, y)$, we have that $\Phi(\cdot, y) + \frac{\rho}{2}\|\cdot\|^2$ is convex for all $y \in \mathbb{R}^n$. We define $\tilde{\Phi}(x, y) = \Phi(x, y) + \frac{\rho}{2}\|x\|^2$ and $\tilde{\psi}(x, y) = \psi(x, y) + \frac{\rho}{2}\|x\|^2$ for $(x, y) \in \mathbb{R}^d \times \mathbb{R}^n$ as well as

$$\tilde{\varphi}(x) = \max_{y \in \mathbb{R}^n} \tilde{\psi}(x, y) = \varphi(x) + \frac{\rho}{2}\|x\|^2.$$

Notice that $\tilde{\psi}(x, \cdot)$ is concave for any $x \in \mathbb{R}^d$ and $\tilde{\psi}(\cdot, y)$ is convex for any $y \in \mathbb{R}^n$. Thus, the function $\tilde{\varphi}$ is convex and $\operatorname{dom}\tilde{\varphi} = \operatorname{dom}\varphi = \mathbb{R}^d$. Therefore φ is continuous, which implies that $\partial\varphi(x) \neq \emptyset$ for any $x \in \mathbb{R}^d$. Let $x \in \mathbb{R}^d$, $y \in Y(x)$ and $v \in \mathbb{R}^d$. For any $\alpha > 0$ it holds

$$\frac{\tilde{\varphi}(x + \alpha v) - \tilde{\varphi}(x)}{\alpha} \geq \frac{\tilde{\psi}(x + \alpha v, y) - \tilde{\psi}(x, y)}{\alpha} = \frac{\tilde{\Phi}(x + \alpha v, y) - \tilde{\Phi}(x, y)}{\alpha},$$

thus

$$\tilde{\varphi}'(x; v) = \inf_{\alpha > 0} \frac{\tilde{\varphi}(x + \alpha v) - \tilde{\varphi}(x)}{\alpha} \geq \inf_{\alpha > 0} \frac{\tilde{\Phi}(x + \alpha v, y) - \tilde{\Phi}(x, y)}{\alpha} = [\tilde{\Phi}(\cdot, y)]'(x; v),$$

where $[\Phi(\cdot, y)]'(x; v)$ denotes the directional derivative of Φ in the first component at x in the direction v. In conclusion,

$$\tilde{\varphi}'(x; v) \geq \sup_{y \in Y(x)} [\tilde{\Phi}(\cdot, y)]'(x; v) \quad \forall v \in \mathbb{R}^d \tag{6.9}$$

and for $y \in Y(x)$ we therefore conclude

$$\partial[\tilde{\Phi}(\cdot, y)](x) \subseteq \partial\tilde{\varphi}(x).$$

The first statement is obtained by subtracting ρx on both sides of the inclusion. □

Lemma 6.3.6 (Lipschitz continuity of the max function)**.** *The Lipschitz continuity of Φ in its first component implies that φ is Lipschitz as well with the same constant.*

Proof. Let $x, x' \in \mathbb{R}^d$ and $y^* \in Y(x)$. On the one hand

$$\begin{aligned}\varphi(x) - \varphi(x') &= \Phi(x, y^*) - h(y^*) - \varphi(x') \\ &\leq \Phi(x, y^*) - h(y^*) - \Phi(x', y^*) + h(y^*) \\ &\leq L\|x - x'\|.\end{aligned}$$

The reverse direction $\varphi(x') - \varphi(x) \leq L\|x - x'\|$ follows analogously. □

6.3.3 Deterministic setting

For initial values $(x_0, y_0) \in \operatorname{dom} f \times \operatorname{dom} h$ the deterministic version of alternating GDA reads as

$$(\forall k \geq 0) \left\lfloor \begin{array}{l} x_{k+1} = \operatorname{prox}_{\eta_x f}(x_k - \eta_x g_k) \\ y_{k+1} = \operatorname{prox}_{\eta_y h}(y_k + \eta_y \nabla_y \Phi(x_{k+1}, y_k)), \end{array} \right. \tag{6.10}$$

for $g_k \in \partial[\Phi(\cdot, y_k)](x_k)$.

Theorem 6.3.7. *Let Assumption 6.3.1, 6.3.2, 6.3.3 and 6.3.4 hold true. The iterates generated by* (6.10) *with the stepsizes $\eta_x = \mathcal{O}(\epsilon^4) < 1/2\rho$, $\eta_y = 1/L_{\nabla\Phi}$ and $\lambda = 1/2\rho$ fulfill*

$$\begin{aligned}&\min_{0 \leq j \leq K-1} \|\nabla g_\lambda(x_j)\|^2 \\ &\qquad \leq 2\epsilon^{-4}\frac{\Delta^*}{K} + 4\rho\epsilon^2\left(L(L + L_f) + L_{\nabla\Phi}C_h^2\right) + 4\rho\frac{\varphi(x_0) - \psi(x_0, y_0)}{K} + 8\epsilon^4\rho L^2,\end{aligned}$$

for $K \geq 1$, where $\Delta^* := g(x_0) - \inf_{x \in \mathbb{R}^d} g(x)$. Therefore, in order to drive the right-hand side to $\mathcal{O}(\epsilon^2)$ and thus to ensure that we visit an ϵ-stationary point, at most $K = \mathcal{O}(\epsilon^{-6})$ iterations are required.

Similarly to the proofs in [35,60] and others, the main descent statement makes use of the quantity $\operatorname{prox}_{\lambda g}(x_k)$ for a $\lambda > 0$. This is somewhat surprising as this point does not appear in the algorithm and can in general not be computed.

But first, we need to establish the fact that $\hat{x}_k := \operatorname{prox}_{\lambda g}(x_k)$ can also be written as the proximal operator of f evaluated at an auxiliary point.

Lemma 6.3.8. *For any $\lambda \in (0, \rho^{-1})$ and all $k \geq 0$ the point $\hat{x}_k := \operatorname{prox}_{\lambda g}(x_k)$ can also be written as*

$$\hat{x}_k = \operatorname{prox}_{\eta_x f}\left(\eta_x \lambda^{-1} x_k - \eta_x v_k + (1 - \eta_x \lambda^{-1})\hat{x}_k\right)$$

for some $v_k \in \partial\varphi(\hat{x}_k)$.

Proof. Let $k \geq 0$ be arbitrary but fixed and recall that $g = f + \varphi$. By the definition of $\hat{x}_k$ we have that

$$0 \in \partial g(\hat{x}_k) + \frac{1}{\lambda}(\hat{x}_k - x_k) = \partial(\varphi + f)(\hat{x}_k) + \frac{1}{\lambda}(\hat{x}_k - x_k).$$

We can estimate through the continuity of φ and subdifferential calculus

$$\frac{1}{\lambda}(x_k - \hat{x}_k) \in \partial(\varphi + f)(\hat{x}_k) \subseteq \partial\varphi(\hat{x}_k) + \partial f(\hat{x}_k).$$

Thus, there exists $v_k \in \partial\varphi(\hat{x}_k)$ such that

$$\frac{1}{\lambda}(x_k - \hat{x}_k) \in v_k + \partial f(\hat{x}_k).$$

Also,

$$\begin{aligned}\eta_x \frac{1}{\lambda}(x_k - \hat{x}_k) \in \eta_x \partial f(\hat{x}_k) + \eta_x v_k &\Leftrightarrow \eta_x \lambda^{-1} x_k - \eta_x v_k + (1 - \eta_x \lambda^{-1})\hat{x}_k \in \hat{x}_k + \eta_x \partial f(\hat{x}_k) \\ &\Leftrightarrow \hat{x}_k = \operatorname{prox}_{\eta_x f}\left(\eta_x \lambda^{-1} x_k - \eta_x v_k + (1 - \eta_x \lambda^{-1})\hat{x}_k\right).\end{aligned}$$

□

With the previous lemma in place we can now turn our attention to the first step of the actual convergence proof.

Lemma 6.3.9. *With* $\lambda = 1/2\rho$ *and* $\eta_x \in [0, \lambda]$ *we have for all* $k \geq 0$ *that*

$$g_\lambda(x_{k+1}) \leq g_\lambda(x_k) + 2\rho\eta_x \Delta_k - \frac{1}{2}\eta_x \|\nabla g_\lambda(x_k)\|^2 + 4\rho\eta_x^2 L^2,$$

where $\Delta_k := \varphi(x_k) - \psi(x_k, y_k) \geq 0$.

Proof. Let $k \geq 0$ be fixed. As in the previous lemma we denote $\hat{x}_k := \operatorname{prox}_{\lambda g}(x_k)$. From the definition of the Moreau envelope we have that

$$g_\lambda(x_{k+1}) = \min_{x \in \mathbb{R}^d} \left\{ g(x) + \frac{1}{2\lambda}\|x - x_{k+1}\|^2 \right\} \leq g(\hat{x}_k) + \frac{1}{2\lambda}\|\hat{x}_k - x_{k+1}\|^2. \tag{6.11}$$

Let now $v_k \in \partial\varphi(\hat{x}_k)$ as in Lemma 6.3.8. We successively deduce

$$\begin{aligned}&\|\hat{x}_k - x_{k+1}\|^2 \\ &\quad = \|\operatorname{prox}_{\eta_x f}\left(\eta_x \lambda^{-1} x_k - \eta_x v_k + (1 - \eta_x \lambda^{-1})\hat{x}_k\right) - \operatorname{prox}_{\eta_x f}(x_k - \eta_x g_k)\|^2 && (6.12) \\ &\quad \leq \|(1 - \eta_x \lambda^{-1})(\hat{x}_k - x_k) + \eta_x (g_k - v_k)\|^2 && (6.13) \\ &\quad = (1 - \eta_x \lambda^{-1})^2 \|\hat{x}_k - x_k\|^2 + 2\eta_x (1 - \eta_x \lambda^{-1})\langle g_k - v_k, \hat{x}_k - x_k \rangle + \eta_x^2 \|g_k - v_k\|^2 \\ &\quad \leq (1 - \eta_x \lambda^{-1})^2 \|\hat{x}_k - x_k\|^2 + 2\eta_x (1 - \eta_x \lambda^{-1})\langle g_k - v_k, \hat{x}_k - x_k \rangle + 4\eta_x^2 L^2 && (6.14)\end{aligned}$$

where (6.12) uses Lemma 6.3.8 and the definition of x_{k+1}, inequality (6.13) holds because of the nonexpansiveness of the proximal operator, and (6.14) follows from the Lipschitz continuity of Φ and φ (see Lemma 6.3.6) and the fact that Lipschitz continuity implies bounded subgradients. We are left with estimating the inner product in the above inequality and we do so by splitting it into two: first of all, from the weak convexity of Φ in x we have that

$$\begin{aligned}\langle g_k, \hat{x}_k - x_k\rangle &\leq \Phi(\hat{x}_k, y_k) - \Phi(x_k, y_k) + \frac{\rho}{2}\|\hat{x}_k - x_k\|^2 \\ &= \Phi(\hat{x}_k, y_k) - h(y_k) - (\Phi(x_k, y_k) - h(y_k)) + \frac{\rho}{2}\|\hat{x}_k - x_k\|^2 \\ &\leq \varphi(\hat{x}_k) - \psi(x_k, y_k) + \frac{\rho}{2}\|\hat{x}_k - x_k\|^2.\end{aligned}$$

Secondly, by the ρ-weak convexity of φ

$$-\langle v_k, \hat{x}_k - x_k\rangle \leq \varphi(x_k) - \varphi(\hat{x}_k) + \frac{\rho}{2}\|\hat{x}_k - x_k\|^2.$$

Combining the last two inequalities we get that

$$\langle g_k - v_k, \hat{x}_k - x_k\rangle \leq \varphi(x_k) - \psi(x_k, y_k) + \rho\|\hat{x}_k - x_k\|^2. \tag{6.15}$$

Plugging (6.15) into (6.14) we deduce

$$\begin{aligned}&\|\hat{x}_k - x_{k+1}\|^2 \\ &\quad\leq \underbrace{\left[(1 - \eta_x\lambda^{-1})^2 + 2\eta_x(1 - \eta_x\lambda^{-1})\rho\right]}_{=(*)}\|\hat{x}_k - x_k\|^2 + 2\eta_x\Delta_k + 4\eta_x^2L^2.\end{aligned} \tag{6.16}$$

Now note that

$$\begin{aligned}(*) &= 1 - 2\eta_x\lambda^{-1} + \eta_x^2\lambda^{-2} + 2\eta_x\rho - 2\eta_x^2\lambda^{-1}\rho \\ &= 1 - 4\eta_x\rho + 4\eta_x^2\rho^2 + 2\eta_x\rho - 4\eta_x^2\rho^2 \\ &= 1 - 2\eta_x\rho.\end{aligned} \tag{6.17}$$

Combining (6.11), (6.16) and (6.17) we deduce,

$$\begin{aligned}g_\lambda(x_{k+1}) &\leq g(\hat{x}_k) + \frac{1}{2\lambda}\left(\|\hat{x}_k - x_k\|^2 + 2\eta_x\Delta_k - 2\eta_x\rho\|\hat{x}_k - x_k\|^2 + 4\eta_x^2L^2\right) \\ &= g_\lambda(x_k) + 2\rho\eta_x\Delta_k - \frac{1}{2}\eta_x\|\nabla g_\lambda(x_k)\|^2 + 4\rho\eta_x^2L^2,\end{aligned}$$

where we used that $\lambda = 1/2\rho$. □

Naturally, we want to telescope the inequality established by the previous lemma. We are left with estimating Δ_k, preferably even in a summable way. But first we need the following technical, yet standard lemma, estimating the amount of increase obtained by a single iteration of gradient ascent.

Lemma 6.3.10. *It holds for all $y \in \mathbb{R}^n$ and $k \geq 0$ that*

$$\psi(x_{k+1}, y) - \psi(x_{k+1}, y_{k+1}) \leq \frac{1}{2\eta_y}\Big(\|y - y_k\|^2 - \|y - y_{k+1}\|^2\Big). \tag{6.18}$$

Proof. By the definition of y_{k+1} we have that

$$y_{k+1} = \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \left\{ h(y) + \Phi(x_{k+1}, y_k) - \langle \nabla_y \Phi(x_{k+1}, y_k), y - y_k \rangle + \frac{1}{2\eta_y} \|y - y_k\|^2 \right\}.$$

Let now $y \in \mathbb{R}^n$ be arbitrary but fixed. Since y_{k+1} minimizes a $1/\eta_y$-strongly convex function,

$$\begin{aligned} &h(y_{k+1}) + \Phi(x_{k+1}, y_k) - \langle \nabla_y \Phi(x_{k+1}, y_k), y_{k+1} - y_k \rangle + \frac{1}{2\eta_y} \|y_{k+1} - y_k\|^2 + \frac{1}{2\eta_y} \|y - y_{k+1}\|^2 \\ &\quad \leq h(y) + \Phi(x_{k+1}, y_k) - \langle \nabla_y \Phi(x_{k+1}, y_k), y - y_k \rangle + \frac{1}{2\eta_y} \|y - y_k\|^2. \end{aligned} \tag{6.19}$$

From the descent lemma (in ascent form) and the fact that $1/\eta_y = L_{\nabla \Phi}$ we have that

$$\Phi(x_{k+1}, y_k) + \langle \nabla_y \Phi(x_{k+1}, y_k), y_{k+1} - y_k \rangle - \frac{1}{2\eta_y} \|y_{k+1} - y_k\|^2 \leq \Phi(x_{k+1}, y_{k+1}). \tag{6.20}$$

From the concavity of Φ in its second component we get that

$$\Phi(x_{k+1}, y) \leq \Phi(x_{k+1}, y_k) + \langle \nabla_y \Phi(x_{k+1}, y_k), y - y_k \rangle. \tag{6.21}$$

By plugging (6.21) and (6.20) into (6.19) we deduce

$$\Phi(x_{k+1}, y) - h(y) + \frac{1}{2\eta_y} \|y - y_{k+1}\|^2 \leq \Phi(x_{k+1}, y_{k+1}) - h(y_{k+1}) + \frac{1}{2\eta_y} \|y - y_k\|^2.$$

The statement of the lemma is obtained by rearranging the above inequality. □

We can now use the previous lemma to estimate Δ_k. Recall also that y_k^* denotes a maximizer of $\psi(x_k, \cdot)$ for all $k \geq 0$.

Lemma 6.3.11. *We have that for all* $1 \leq m \leq k$,

$$\Delta_k \leq 2\eta_x L(L + L_f)(k - m) + \frac{1}{2\eta_y} \Big(\|y_{k-1} - y_m^*\|^2 - \|y_k - y_m^*\|^2 \Big). \tag{6.22}$$

Proof. Let $1 \leq m \leq k$ be fixed. Plugging $y = y_m^*$ into (6.18) we deduce that

$$0 \leq \psi(x_k, y_k) - \psi(x_k, y_m^*) + \frac{1}{2\eta_y} \Big(\|y_m^* - y_{k-1}\|^2 - \|y_m^* - y_k\|^2 \Big). \tag{6.23}$$

Starting from the definition of Δ_k, we add (6.23) to obtain

$$\begin{aligned} \Delta_k &= \psi(x_k, y_k^*) - \psi(x_k, y_k) \\ &\leq \psi(x_k, y_k^*) - \psi(x_k, y_m^*) + \frac{1}{2\eta_y} \Big(\|y_m^* - y_{k-1}\|^2 - \|y_m^* - y_k\|^2 \Big). \end{aligned} \tag{6.24}$$

Due to the Lipschitz continuity of Φ, terms which only differ in their first argument will be easy to estimate. Therefore, we insert and subtract $\Phi(x_m, y^*_{k+1})$ to deduce

$$\begin{aligned}
&\psi(x_k, y^*_k) - \psi(x_k, y^*_m) \\
&\quad = \Phi(x_k, y^*_k) - h(y^*_k) - \Phi(x_k, y^*_m) + h(y^*_m) \\
&\quad = \Phi(x_k, y^*_k) - \Phi(x_m, y^*_k) + \Phi(x_m, y^*_k) - h(y^*_k) - \Phi(x_k, y^*_m) + h(y^*_m) \\
&\quad \leq \Phi(x_k, y^*_k) - \Phi(x_m, y^*_k) + \Phi(x_m, y^*_m) - h(y^*_m) - \Phi(x_k, y^*_m) + h(y^*_m) \\
&\quad = \Phi(x_k, y^*_k) - \Phi(x_m, y^*_k) + \Phi(x_m, y^*_m) - \Phi(x_k, y^*_m).
\end{aligned} \tag{6.25}$$

We estimate the above expression for $k > m$ by making use of the Lipschitz continuity of $\Phi(\cdot, y)$ and (6.7) deducing

$$\begin{aligned}
&\Phi(x_k, y^*_k) - \Phi(x_m, y^*_k) \\
&\quad \leq L\|x_k - x_m\| \leq L \sum_{l=m}^{k-1} \|x_{l+1} - x_l\| \\
&\quad \leq L \sum_{l=m}^{k-1} \Big(\|\operatorname{prox}_{\eta_x f}(x_l - \eta_x g_l) - \operatorname{prox}_{\eta_x f}(x_l)\| + \|\operatorname{prox}_{\eta_x f}(x_l) - x_l\| \Big) \\
&\quad \leq \eta_x L(L + L_f)(k - m).
\end{aligned} \tag{6.26}$$

For $k = m$ the inequality follows trivially. Analogously, we deduce

$$\Phi(x_m, y^*_m) - \Phi(x_k, y^*_m) \leq \eta_x L(L + L_f)(k - m). \tag{6.27}$$

Plugging (6.25), (6.26) and (6.27) into (6.24) gives the statement of the lemma. □

In order to estimate the summation of Δ_k we will use a trick to sum over it in blocks, where the size B of these blocks will depend on the total number of iterations K. Note that w.l.o.g. we assume that the block size $B \leq K$ divides K without remainder.

Lemma 6.3.12. *It holds that for all $K \geq 1$*

$$\frac{1}{K} \sum_{k=0}^{K-1} \Delta_k \leq \eta_x L(L + L_f)B + \frac{L_{\nabla\Phi} C_h^2}{2B} + \frac{\Delta_0}{K}. \tag{6.28}$$

Proof. By splitting the summation into blocks we get that

$$\sum_{k=0}^{K-1} \Delta_k = \sum_{j=0}^{K/B-1} \sum_{k=jB}^{(j+1)B-1} \Delta_k. \tag{6.29}$$

By using (6.22) from Lemma 6.3.11 with $m = 1$ and the fact that $\sum_{k=1}^{B-1} k \leq B^2/2$ we get that

$$\begin{aligned}
\sum_{k=0}^{B-1} \Delta_k &\leq \Delta_0 + \eta_x L(L + L_f)B^2 + \frac{1}{2\eta_y}\|y_0 - y^*_1\|^2 \\
&\leq \Delta_0 + \eta_x L(L + L_f)B^2 + \frac{1}{2\eta_y} C_h^2,
\end{aligned} \tag{6.30}$$

where C_h was defined in Assumption 6.3.4 and denotes the diameter of $\operatorname{dom} h$. Analogously, for $j > 0$ and $m = jB$ we have that

$$\begin{aligned}\sum_{k=jB}^{(j+1)B-1} \Delta_k &\leq \eta_x L(L+L_f)B^2 + \frac{1}{2\eta_y}\|y_{jB-1} - y^*_{jB}\|^2 \\ &\leq \eta_x L(L+L_f)B^2 + \frac{1}{2\eta_y} C_h^2.\end{aligned} \tag{6.31}$$

Plugging (6.30) and (6.31) into (6.29) gives

$$\frac{1}{K}\sum_{k=0}^{K-1} \Delta_k \leq \eta_x L(L+L_f)B + \frac{1}{2\eta_y B} C_h^2 + \frac{\Delta_0}{K}.$$

The desired statement is obtained by using the stepsize $\eta_y = {}^1\!/\!{}_{L_{\nabla\Phi}}$. □

Proof of Theorem 6.3.7. From Lemma 6.3.9 we deduce by summing up

$$g_\lambda(x_K) \leq g_\lambda(x_0) + 2\eta_x\rho\sum_{k=0}^{K-1}\Delta_k - \frac{1}{2}\eta_x\sum_{k=0}^{K-1}\|\nabla g_\lambda(x_k)\|^2 + 4K\rho\eta_x^2L^2.$$

Next, we divide by K and obtain that

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla g_\lambda(x_k)\|^2 \leq 2\frac{\Delta^*}{\eta_x K} + \frac{4\rho}{K}\sum_{k=0}^{K-1}\Delta_k + 8\rho\eta_x L^2.$$

Now, we plug in (6.28) to deduce that

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla g_\lambda(x_k)\|^2 \leq 2\frac{\Delta^*}{\eta_x K} + 4\rho\Big(\eta_x L(L+L_f)B + \frac{L_{\nabla\Phi}C_h^2}{2B}\Big) + \frac{4\rho\Delta_0}{K} + 8\eta_x\rho L^2.$$

With $B = {}^1\!/\!{}_{\sqrt{\eta_x}}$, we have that

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla g_\lambda(x_k)\|^2 \leq 2\frac{\Delta^*}{\eta_x K} + \rho\sqrt{\eta_x}\Big(4L(L+L_f) + 2L_{\nabla\Phi}C_h^2\Big) + \frac{4\rho\Delta_0}{K} + 8\eta_x\rho L^2.$$

We obtain the statement of the theorem by plugging in $\eta_x = \mathcal{O}(\epsilon^4)$. □

6.3.4 Stochastic setting

For initial values $(x_0, y_0) \in \operatorname{dom} f \times \operatorname{dom} h$ the stochastic version of alternating GDA is given by

$$(\forall k \geq 0) \quad \left\lfloor \begin{array}{l} x_{k+1} = \operatorname{prox}_{\eta_x f}\left(x_k - \eta_x g_k^\xi\right) \\ y_{k+1} = \operatorname{prox}_{\eta_y h}\left(y_k + \eta_y \nabla_y \Phi(x_{k+1}, y_k; \zeta_k)\right), \end{array}\right. \tag{6.32}$$

for $g_k^\xi \in \partial[\Phi(\cdot, y_k; \xi_k)](x_k)$ for $\xi_k, \zeta_k \sim \mathcal{D}$ independent from all previous iterates.

Theorem 6.3.13. *Let in addition to the assumptions of Theorem 6.3.7 also Assumption 6.2.1 and 6.2.2 hold true. The iterates generated by* (6.32) *with* $\eta_y = \mathcal{O}(\epsilon^2) \leq 1/2L_{\nabla\Phi}$, $\eta_x = \mathcal{O}(\epsilon^6) < 1/2\rho$ *and* $\lambda = 1/2\rho$ *fulfill*

$$\begin{aligned}&\min_{0\leq j\leq K-1} \mathbb{E}\big[\|\nabla g_\lambda(x_k)\|^2\big]\\ &\qquad\leq \frac{2\Delta^*}{K}\epsilon^{-6} + 4\rho\epsilon^2(L(L+L_f+\sigma)+C_h^2+\sigma^2) + 4\rho\frac{\varphi(x_0)-\psi(x_0,y_0)}{K} + 8\epsilon^6\rho(L^2+\sigma^2),\end{aligned}$$

for $K \geq 1$, *where* $\Delta^* = g(x_0) - \inf_{x\in\mathbb{R}^d} g(x)$. *Therefore, in order to drive the right hand side to* $\mathcal{O}(\epsilon^2)$ *and thus to ensure that we visit an* ϵ*-stationary point, at most* $K = \mathcal{O}(\epsilon^{-8})$ *iterations are required.*

The proof proceeds along the same lines of the deterministic case. Similarly we show an adapted version of Lemma 6.3.9.

Lemma 6.3.14. *With* $\lambda = 1/2\rho$ *we have for all* $k \geq 0$ *that*

$$\mathbb{E}[g_\lambda(x_{k+1})] \leq \mathbb{E}[g_\lambda(x_k)] + 2\rho\eta_x\hat{\Delta}_k - \frac{\eta_x}{2}\mathbb{E}\big[\|\nabla g_\lambda(x_k)\|^2\big] + 4\rho\eta_x^2(L^2+\sigma^2)$$

where $\hat{\Delta}_k := \mathbb{E}[\varphi(x_k) - \psi(x_k,y_k)]$.

Proof. Let $k \geq 0$ be arbitrary but fixed. Note that it follows easily from (6.6) that

$$\mathbb{E}\Big[\|g_k^\xi\|^2\Big] \leq \mathbb{E}\big[\|g_k\|^2\big] + \sigma^2 \leq L^2 + \sigma^2, \tag{6.33}$$

where $\mathbb{E}\Big[g_k^\xi\Big] = g_k \in \partial_x\Phi(x_k,y_k)$. From the definition of the Moreau envelope we deduce that

$$\mathbb{E}[g_\lambda(x_{k+1})] \leq \mathbb{E}[g(\hat{x}_k)] + \frac{1}{2\lambda}\mathbb{E}\big[\|\hat{x}_k - x_{k+1}\|^2\big]. \tag{6.34}$$

Similarly to Lemma 6.3.9 we deduce that for $v_k \in \partial\varphi(\hat{x}_k)$ (as given in Lemma 6.3.8)

$$\begin{aligned}&\|\hat{x}_k - x_{k+1}\|^2\\ &\quad= \|\mathrm{prox}_{\eta_x f}\left(\eta_x\lambda^{-1}x_k - \eta_x v_k + (1-\eta_x\lambda^{-1})\hat{x}_k\right) - \mathrm{prox}_{\eta_x f}\left(x_k - \eta_x g_k^\xi\right)\|^2\\ &\quad\leq \|(1-\eta_x\lambda^{-1})(\hat{x}_k - x_k) + \eta_x(g_k^\xi - v_k)\|^2\\ &\quad= (1-\eta_x\lambda^{-1})^2\|\hat{x}_k - x_k\|^2 + 2\eta_x(1-\eta_x\lambda^{-1})\langle g_k^\xi - v_k, \hat{x}_k - x_k\rangle + \eta_x^2\|g_k^\xi - v_k\|^2.\end{aligned}$$

By applying the conditional expectation $\mathbb{E}[\cdot\,|\,x_k,y_k]$, then the unconditional one and using (6.33), we get that

$$\begin{aligned}&\mathbb{E}\big[\|\hat{x}_k - x_{k+1}\|^2\big]\\ &\qquad\leq \mathbb{E}\big[\|\hat{x}_k - x_k\|^2\big] + 2\eta_x(1-\eta_x\lambda^{-1})\mathbb{E}[\langle g_k - v_k, \hat{x}_k - x_k\rangle] + 4\eta_x^2(L^2+\sigma^2).\end{aligned}$$

where $g_k = \mathbb{E}\Big[g_k^\xi\Big]$. Lastly, we combine the above inequality with (6.34) and the estimate for the inner product (6.15) as in Lemma 6.3.9 to deduce the statement of the lemma. □

Next, we discuss the stochastic version of Lemma 6.3.10. It is clear that we cannot expect the same amount of function value increase by a single iteration of gradient ascent if we do not use the exact gradient.

Lemma 6.3.15. *With $\eta_y \leq {}^1\!/_{2L_{\nabla\Phi}}$ we have for all $k \geq 0$ and all $y \in \mathbb{R}^n$*

$$\mathbb{E}[\psi(x_{k+1}, y) - \psi(x_{k+1}, y_{k+1})] \leq \frac{1}{2\eta_y}\Big(\mathbb{E}\big[\|y - y_k\|^2\big] - \mathbb{E}\big[\|y - y_{k+1}\|^2\big]\Big) + \eta_y\sigma^2. \quad (6.35)$$

Proof. Let $k \geq 0$ and $y \in \mathbb{R}^n$ be arbitrary but fixed. By the definition of y_{k+1} we have that

$$y_{k+1} = \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \Big\{h(y) - \langle \nabla_y \Phi(x_{k+1}, y_k; \zeta_k), y - y_k\rangle + \frac{1}{2\eta_y}\|y - y_k\|^2\Big\}.$$

Therefore, as in Lemma 6.3.10, we deduce that

$$\begin{aligned} h(y_{k+1}) - \langle \nabla_y \Phi(x_{k+1}, y_k; \zeta_k), y_{k+1} - y_k\rangle + \frac{1}{2\eta_y}\|y_{k+1} - y_k\|^2 + \frac{1}{2\eta_y}\|y - y_{k+1}\|^2 \\ \leq h(y) - \langle \nabla_y \Phi(x_{k+1}, y_k; \zeta_k), y - y_k\rangle + \frac{1}{2\eta_y}\|y - y_k\|^2. \end{aligned}$$

The term $\langle \nabla_y \Phi(x_{k+1}, y_k; \zeta_k), y_{k+1} - y_k\rangle$ is problematic, because the right hand side of the inner product is not measurable with respect to the sigma algebra generated by (x_{k+1}, y_k), so we insert and subtract $\nabla_y \Phi(x_{k+1}, y_k)$ to deduce

$$\begin{aligned} &h(y_{k+1}) + \langle \nabla_y \Phi(x_{k+1}, y_k; \zeta_k), y - y_k\rangle + \frac{1}{2\eta_y}\|y_{k+1} - y_k\|^2 + \frac{1}{2\eta_y}\|y - y_{k+1}\|^2 \\ &\quad \leq h(y) + \langle \nabla_y \Phi(x_{k+1}, y_k), y_{k+1} - y_k\rangle \\ &\qquad + \langle \nabla_y \Phi(x_{k+1}, y_k; \zeta_k) - \nabla_y \Phi(x_{k+1}, y_k), y_{k+1} - y_k\rangle + \frac{1}{2\eta_y}\|y - y_k\|^2. \end{aligned}$$

Now, using Young's inequality we get that

$$\begin{aligned} &\langle \nabla_y \Phi(x_{k+1}, y_k; \zeta_k) - \nabla_y \Phi(x_{k+1}, y_k), y_{k+1} - y_k\rangle \\ &\qquad \leq \eta_y\|\nabla_y \Phi(x_{k+1}, y_k; \zeta_k) - \nabla_y \Phi(x_{k+1}, y_k)\|^2 + \frac{1}{4\eta_y}\|y_{k+1} - y_k\|^2. \end{aligned}$$

Combining the above two inequalities and taking the conditional expectation gives

$$\begin{aligned} &\langle \nabla_y \Phi(x_{k+1}, y_k), y - y_k\rangle + \mathbb{E}\left[h(y_{k+1}) + \frac{1}{2\eta_y}\|y - y_{k+1}\|^2 \,\middle|\, x_{k+1}, y_k\right] \\ &\quad \leq h(y) + \mathbb{E}\left[\langle \nabla_y \Phi(x_{k+1}, y_k), y_{k+1} - y_k\rangle - \frac{1}{4\eta_y}\|y_{k+1} - y_k\|^2 \,\middle|\, x_{k+1}, y_k\right] \\ &\qquad + \eta_y\mathbb{E}\big[\|\nabla_y \Phi(x_{k+1}, y_k; \zeta_k) - \nabla_y \Phi(x_{k+1}, y_k)\|^2 \,\big|\, x_{k+1}, y_k\big] + \frac{1}{2\eta_y}\|y - y_k\|^2. \end{aligned}$$

And now the unconditional expectation together with the bounded variance assumption (6.5)

$$\begin{aligned}\mathbb{E}[h(y_{k+1}) + \langle \nabla_y \Phi(x_{k+1}, y_k), y - y_k\rangle] + \frac{1}{2\eta_y}\mathbb{E}\big[\|y - y_{k+1}\|^2\big] \\ \leq \mathbb{E}[h(y) + \langle \nabla_y \Phi(x_{k+1}, y_k), y_{k+1} - y_k\rangle] - \frac{1}{4\eta_y}\mathbb{E}\big[\|y_{k+1} - y_k\|^2\big] \\ + \frac{1}{2\eta_y}\mathbb{E}\big[\|y - y_k\|^2\big] + \eta_y \sigma^2.\end{aligned} \tag{6.36}$$

From the descent lemma (in ascent form) and the fact that $\eta_y \leq 1/2L_{\nabla\Phi}$ we have that

$$\Phi(x_{k+1}, y_k) + \langle y_{k+1} - y_k, \nabla_y \Phi(x_{k+1}, y_k)\rangle - \frac{1}{4\eta_y}\|y_{k+1} - y_k\|^2 \leq \Phi(x_{k+1}, y_{k+1}).$$

We plug the above inequality into (6.36) and also make use of the concavity as in (6.21) to deduce the statement of the lemma. □

We can now use the previous lemma to estimate $\hat{\Delta}_k$.

Lemma 6.3.16. *For all $1 \leq m \leq k$, we have that*

$$\hat{\Delta}_k \leq 2\eta_x L(L_f + L + \sigma)(k - m) + \frac{1}{2\eta_y}\Big(\mathbb{E}\big[\|y_{k-1} - y_m^*\|^2\big] - \mathbb{E}\big[\|y_k - y_m^*\|^2\big]\Big) + \eta_y \sigma^2. \tag{6.37}$$

Proof. Let the numbers $1 \leq m \leq k$ be fixed. Plugging in $y = y_m^*$ into (6.35) we deduce that

$$0 \leq \mathbb{E}[\psi(x_k, y_k) - \psi(x_k, y_m^*)] + \frac{1}{2\eta_y}\Big(\mathbb{E}\big[\|y_m^* - y_{k-1}\|^2\big] - \mathbb{E}\big[\|y_m^* - y_k\|^2\big]\Big) + \eta_y \sigma^2. \tag{6.38}$$

Starting from the definition of $\hat{\Delta}_k$, we add (6.38) to obtain

$$\begin{aligned}\hat{\Delta}_k &= \mathbb{E}[\psi(x_k, y_k^*) - \psi(x_k, y_k)] \\ &\leq \mathbb{E}[\psi(x_k, y_k^*) - \psi(x_k, y_m^*)] + \frac{1}{2\eta_y}\Big(\mathbb{E}\big[\|y_m^* - y_{k-1}\|^2\big] - \mathbb{E}\big[\|y_m^* - y_k\|^2\big]\Big) + \eta_y \sigma^2.\end{aligned} \tag{6.39}$$

As in (6.25) we deduce that

$$\psi(x_k, y_k^*) - \psi(x_k, y_m^*) \leq \Phi(x_k, y_k^*) - \Phi(x_m, y_k^*) + \Phi(x_m, y_m^*) - \Phi(x_k, y_m^*).$$

Together with the L-Lipschitz continuity of $\Phi(\cdot, y)$ and (6.7) we estimate for $k > m$ that

$$\begin{aligned}&\mathbb{E}[\Phi(x_k, y_k^*) - \Phi(x_m, y_k^*)] \\ &\quad\leq L\mathbb{E}[\|x_k - x_m\|] \leq L\sum_{l=m}^{k-1}\mathbb{E}[\|x_{l+1} - x_l\|] \\ &\quad\leq L\sum_{l=m}^{k-1}\Big(\mathbb{E}\Big[\|\text{prox}_{\eta_x f}\Big(x_l - \eta_x g_l^{\xi}\Big) - \text{prox}_{\eta_x f}(x_l)\|\Big] + \mathbb{E}\big[\|\text{prox}_{\eta_x f}(x_l) - x_l\|\big]\Big) \\ &\quad\leq \eta_x L\Big(L_f + \sqrt{L^2 + \sigma^2}\Big)(k - m).\end{aligned}$$

For $k = m$ the statement follows trivially. Analogously,

$$\mathbb{E}[\Phi(x_m, y_m^*) - \Phi(x_k, y_m^*)] \le L\mathbb{E}[\|x_k - x_m\|] \le \eta_x L\Big(L_f + \sqrt{L^2 + \sigma^2}\Big)(k - m).$$

Plugging all of these into (6.39) gives the statement of the lemma. □

In order to estimate the summation of $\hat{\Delta}_k$ we will use the same trick as in the deterministic setting and sum over it in blocks, where the size B of these blocks will divide the total number of iterations K.

Lemma 6.3.17. *We have that for all $K \ge 1$*

$$\frac{1}{K}\sum_{k=0}^{K-1} \hat{\Delta}_k \le \eta_x L(L + L_f + \sigma)B + \frac{C_h^2}{2\eta_y B} + \eta_y \sigma^2 + \frac{\Delta_0}{K}. \tag{6.40}$$

Proof. By using (6.37) from Lemma 6.3.16 with $m = 1$ and the fact that $\sum_{k=1}^{B-1} k \le {}^{B^2}/_2$ we get that

$$\begin{aligned}\sum_{k=0}^{B-1} \hat{\Delta}_k &\le \Delta_0 + \eta_x L(L + L_f + \sigma)B^2 + \frac{1}{2\eta_y}\mathbb{E}\big[\|y_0 - y_1^*\|^2\big] + B\eta_y\sigma^2 \\ &\le \Delta_0 + \eta_x L(L + L_f + \sigma)B^2 + \frac{1}{2\eta_y}C_h^2 + B\eta_y\sigma^2,\end{aligned} \tag{6.41}$$

where C_h denotes the diameter of $\operatorname{dom} h$, see Assumption 6.3.4. Analogously, for $j > 0$ and $m = jB$ we have that

$$\begin{aligned}\sum_{k=jB}^{(j+1)B-1} \hat{\Delta}_k &\le \eta_x L(L + L_f + \sigma)B^2 + \frac{1}{2\eta_y}\mathbb{E}\big[\|y_{jB-1} - y_{jB}^*\|^2\big] + B\eta_y\sigma^2 \\ &\le \eta_x L(L + L_f + \sigma)B^2 + \frac{1}{2\eta_y}C_h^2 + B\eta_y\sigma^2,\end{aligned} \tag{6.42}$$

Plugging (6.41) and (6.42) into (6.29) gives the statement of the lemma. □

Now we can prove the convergence result for the stochastic algorithm.

Proof of Theorem 6.3.13. We sum up the inequality of Lemma 6.3.14 to deduce that

$$\mathbb{E}[g_\lambda(x_K)] \le \mathbb{E}[g_\lambda(x_0)] + 2\eta_x\rho\sum_{k=0}^{K-1}\hat{\Delta}_k - \frac{\eta_x}{2}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla g_\lambda(x_k)\|^2\big] + 4K\rho\eta_x^2(L^2 + \sigma^2).$$

Thus, by dividing by K and η_x yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[\|\nabla g_\lambda(x_k)\|^2\big] \le \frac{2\Delta^*}{\eta_x K} + \frac{4\rho}{K}\sum_{k=0}^{K-1}\hat{\Delta}_k + 8\rho\eta_x(L^2 + \sigma^2).$$

Now we plug in (6.40) to obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big[\|\nabla g_\lambda(x_k)\|^2\big] \\ \leq \frac{2\Delta^*}{\eta_x K} + 4\rho\Big(\eta_x L(L+L_f+\sigma)B + \frac{C_h^2}{2\eta_y B} + \eta_y\sigma^2\Big) + 4\rho\frac{\Delta_0}{K} + 8\rho\eta_x(L^2+\sigma^2).$$

With the stepsize $\eta_y = \epsilon^2$, $\eta_x = \epsilon^6$ and $B = \epsilon^{-4}$ we have that

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla g_\lambda(x_k)\|^2 \leq \frac{2\Delta^*}{\epsilon^6 K} + 2\rho\epsilon^2(2L(L+L_f+\sigma) + C_h^2 + 2\sigma^2) + 4\rho\frac{\Delta_0}{K} + 8\rho\epsilon^6(L^2+\sigma^2),$$

which finishes the proof. □

6.4 Nonconvex-strongly concave objective

Instead of the assumptions of Section 6.3 we require only the following ones.

Assumption 6.4.1. *Let Φ be $L_{\nabla\Phi}$-smooth uniformly in both components and concave in the second one. The regularizers f and $-h$ are proper, convex and lower semicontinuous. Additionally, either Φ is μ-strongly concave in the second component, uniformly in the first one, or $-h$ is μ-strongly concave.*

Assumption 6.4.2. *Let $g = \varphi + f$ be lower bounded, i.e. $\inf_{x\in\mathbb{R}^d} g(x) > -\infty$.*

Notation. In Proposition 6.4.4 we will show that under the above assumptions $\varphi = \max_{y\in\mathbb{R}^n}\{\Phi(\cdot,y) - h(y)\}$ is $L_{\nabla\varphi}$-smooth, with $L_{\nabla\varphi} = (1+\kappa)L_{\nabla\Phi}$, for $\kappa := \max\{{}^{L_{\nabla\Phi}}/_{\mu}, 1\}$ denoting the *condition number.* In the setting without regularizers, where the strong concavity arises from Φ it is well known that $\mu \leq L_{\nabla\Phi}$ and therefore $1 \leq {}^{L_{\nabla\Phi}}/_{\mu}$ (which is the standard definition of the condition number). If the strong concavity stems from the regularizers h this is no longer true and ${}^{L_{\nabla\Phi}}/_{\mu}$ might be smaller than 1 which would lead to tedious case distinctions, which is why we adapt the definition of the condition number in order to provide a unified analysis. Additionally, the solution set $Y(x)$ defined in (6.8) consists only of a single element which we will denote by $y^*(x)$. We denote the quantity $\delta_k := \|y_k - y_k^*\|^2$, measuring the distance between the current strategy of the second player and her best response according to the current strategy of the first player.

6.4.1 Properties of the max function

In the following we will show the smoothness of φ, as well as the fact that the solution map fulfills a strong Lipschitz property.

Lemma 6.4.3 (Lipschitz continuity of the solution mapping)**.** *The solution map $y^*: \mathbb{R}^d \to \mathbb{R}^n$ which fulfills $\psi(x, y^*(x)) = \max_{y\in\mathbb{R}^n}\psi(x,y)$ for all $x \in \mathbb{R}^d$ is well defined and κ-Lipschitz where $\kappa = \max\{{}^{L_{\nabla\Phi}}/_{\mu}, 1\}$.*

Proof. Let $x, x' \in \mathbb{R}^d$ be fixed. From the optimality condition we deduce that

$$0 \in \partial h(y^*(x)) - \nabla_y \Phi(x, y^*(x))$$

and

$$\nabla_y \Phi(x', y^*(x')) - \nabla_y \Phi(x, y^*(x')) \in \partial h(y^*(x')) - \nabla_y \Phi(x, y^*(x')).$$

Thus by the strong monotonicity of $\partial h - \nabla_y \Phi(x, \cdot)$ we obtain

$$\begin{aligned}\mu \|y^*(x) - y^*(x')\|^2 &\leq \langle y^*(x) - y^*(x'), \nabla_y \Phi(x, y^*(x')) - \nabla_y \Phi(x', y^*(x'))\rangle \\ &\leq \|y^*(x) - y^*(x')\| \|\nabla_y \Phi(x, y^*(x')) - \nabla_y \Phi(x', y^*(x'))\| \\ &\leq \|y^*(x) - y^*(x')\| L_{\nabla \Phi} \|x - x'\|.\end{aligned}$$

The statement of the lemma follows. □

Proposition 6.4.4 (Smoothness of the max function)**.** *Let Assumption 6.4.1 hold true. Then, φ is smooth and its gradient is given by*

$$\nabla \varphi(x) = \nabla_x \Phi(x, y^*(x))$$

and is therefore $L_{\nabla \Phi}(1 + \kappa)$-Lipschitz.

Proof. Following the notation of Proposition 6.3.5, we introduce $\tilde{\varphi}(x, y) = \varphi(x, y) + (L_{\nabla \Phi}/2)\|x\|^2$, $\tilde{\Phi}(x, y) = \Phi(x, y) + (L_{\nabla \Phi}/2)\|x\|^2$ and $\tilde{\psi}(x, y) = \psi(x, y) + (L_{\nabla \Phi}/2)\|x\|^2$ for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^n$. Let $x, v \in \mathbb{R}^d$, $\alpha_k \downarrow 0$ and $x^k := x + \alpha_k v$ for any $k \geq 0$. Further, let be $y^k = y^*(x^k)$ for any $k \geq 0$. Then, by the Lipschitz continuity of $y^*(\cdot)$, see Lemma 6.4.3, $\lim_{k \to \infty} y^k = y^*(x)$. In addition, for any $v \in \mathbb{R}^d$ and all $k \geq 0$,

$$\begin{aligned}\tilde{\varphi}'(x; v) \leq \frac{\tilde{\varphi}(x + \alpha_k v) - \tilde{\varphi}(x)}{\alpha_k} &\leq \frac{\tilde{\psi}(x + \alpha_k v, y^k) - \tilde{\psi}(x, y^k)}{\alpha_k} \\ &= \frac{\tilde{\Phi}(x + \alpha_k v, y^k) - \tilde{\Phi}(x, y^k)}{\alpha_k} \\ &\leq -[\tilde{\Phi}(\cdot, y^k)]'(x + \alpha_k v; -v) \leq [\tilde{\Phi}(\cdot, y^k)]'(x + \alpha_k v; v).\end{aligned}$$

In other words, for any $v \in \mathbb{R}^d$,

$$\tilde{\varphi}'(x; v) \leq \langle \nabla_x \tilde{\Phi}(x^k, y^k), v\rangle \quad \forall k \geq 0.$$

Since the gradient of $\tilde{\Phi}$ is continuous, this implies by letting $k \to +\infty$ that

$$\tilde{\varphi}'(x; v) \leq \langle \nabla_x \tilde{\Phi}(x, y^*(x)), v\rangle, \quad \forall v \in \mathbb{R}^d,$$

which, together with (6.9), yields

$$\tilde{\varphi}'(x; v) = \langle \nabla_x \Phi(x, y^*(x)), v\rangle \quad \forall v \in \mathbb{R}^d.$$

The fact that the gradient of φ is Lipschitz continuous follows from

$$\begin{aligned}&\|\nabla \varphi(x) - \nabla \varphi(x')\| \\ &\quad \leq \|\nabla_x \Phi(x, y^*(x)) - \nabla_x \Phi(x', y^*(x))\| + \|\nabla_x \Phi(x', y^*(x)) - \nabla_x \Phi(x', y^*(x'))\| \\ &\quad \leq L_{\nabla \Phi}\|x - x'\| + L_{\nabla \Phi}\|y^*(x) - y^*(x')\| \leq (L_{\nabla \Phi} + L_{\nabla \Phi}\kappa)\|x - x'\|,\end{aligned}$$

together with the claimed constant. □

6.4.2 Deterministic setting

For the purpose of this section Algorithm 6.2.3 reads as

$$(\forall k \geq 0) \left\lfloor \begin{array}{l} x_{k+1} = \operatorname{prox}_{\eta_x f}\left(x_k - \eta_x \nabla_x \Phi(x_k, y_k)\right) \\ y_{k+1} = \operatorname{prox}_{\eta_y h}\left(y_k + \eta_y \nabla_y \Phi(x_{k+1}, y_k)\right), \end{array} \right. \tag{6.43}$$

We start with the main convergence result of this section.

Theorem 6.4.5. *Let Assumption 6.4.1 and 6.4.2 hold. The iterates generated by Algorithm 6.2.3 with stepsize $\eta_y = 1/L_{\nabla\Phi}$ and $\eta_x = 1/(3(\kappa+1)^2 L_{\nabla\Phi})$ fulfill*

$$\min_{1 \leq k \leq K} \operatorname{dist}\Big(- \nabla\varphi(x_k), \partial f(x_k)\Big)^2 \leq 6(\kappa+1)^2 L_{\nabla\Phi} \frac{\Delta^*}{K} + 4\frac{L_{\nabla\Phi}^2 \kappa \|y^*(x_0) - y_0\|^2}{K}.$$

for $\Delta^* = g(x_0) - \inf_{x\in\mathbb{R}^d} g(x)$. This means that an ϵ-stationary point is visited in at most $\mathcal{O}(\kappa^2\epsilon^{-2})$ iterations.

Let us start with the first lemma.

Lemma 6.4.6. *There exists a sequence of points $(w_k)_{k\geq 1}$ such that $w_k \in (\partial f + \nabla\varphi)(x_k)$ and its norm can be bounded by*

$$\frac{1}{2}\eta_x \|w_{k+1}\|^2 \leq g(x_k) - g(x_{k+1}) + \frac{1}{2}\Big(L_{\nabla\varphi} + 2L_{\nabla\varphi}^2 \eta_x - \frac{1}{\eta_x}\Big)\|x_k - x_{k+1}\|^2 + \eta_x L_{\nabla\Phi}^2 \delta_k$$

for all $k \geq 0$.

Proof. Let $k \geq 0$ be arbitrary but fixed. From the definition of the proximal operator we deduce that

$$0 \in \partial f(x_{k+1}) + \nabla_x \Phi(x_k, y_k) + \frac{1}{\eta_x}(x_{k+1} - x_k)$$

Thus,

$$w_{k+1} := \frac{1}{\eta_x}(x_k - x_{k+1}) + \nabla\varphi(x_{k+1}) - \nabla_x\Phi(x_k, y_k) \in \partial f(x_{k+1}) + \nabla\varphi(x_{k+1}),$$

as claimed. In order to prove the bound on $\|w_{k+1}\|$ we proceed as follows:

$$\begin{aligned} \|w_{k+1}\|^2 = \eta_x^{-2}\|x_k - x_{k+1}\|^2 &+ 2\eta_x^{-1}\langle x_k - x_{k+1}, \nabla\varphi(x_{k+1}) - \nabla_x\Phi(x_k, y_k)\rangle \\ &+ \|\nabla\varphi(x_{k+1}) - \nabla_x\Phi(x_k, y_k)\|^2. \end{aligned} \tag{6.44}$$

The smoothness of φ implies via the descent lemma that

$$\varphi(x_{k+1}) + \langle \nabla\varphi(x_{k+1}), x_k - x_{k+1}\rangle - \frac{L_{\nabla\varphi}}{2}\|x_{k+1} - x_k\|^2 \leq \varphi(x_k). \tag{6.45}$$

Since the proximal operator minimizes a $1/\eta_x$-strongly convex function we have that

$$\begin{aligned} f(x_{k+1}) + \langle \nabla_x\Phi(x_k, y_k), x_{k+1} - x_k\rangle + \frac{1}{2\eta_x}\|x_{k+1} - x_k\|^2 + \frac{1}{2\eta_x}\|x_{k+1} - x\|^2 \\ \leq f(x) + \langle \nabla_x\Phi(x_k, y_k), x - x_k\rangle + \frac{1}{2\eta_x}\|x - x_k\|^2 \end{aligned}$$

for all $x \in \mathbb{R}^d$. Adding this inequality at $x = x_k$ to (6.45) we deduce that

$$\langle \nabla \varphi(x_{k+1}) - \nabla_x \Phi(x_k, y_k), x_k - x_{k+1} \rangle \leq g(x_k) - g(x_{k+1}) + \frac{1}{2} \left(L_{\nabla \varphi} - \frac{2}{\eta_x} \right) \|x_{k+1} - x_k\|^2. \tag{6.46}$$

Lastly, by the Young inequality

$$\begin{aligned} \|\nabla \varphi(x_{k+1}) - \nabla_x \Phi(x_k, y_k)\|^2 &= \|\nabla \varphi(x_{k+1}) - \nabla \varphi(x_k) + \nabla \varphi(x_k) - \nabla_x \Phi(x_k, y_k)\|^2 \\ &\leq 2L_{\nabla \varphi}^2 \|x_{k+1} - x_k\|^2 + 2L_{\nabla \Phi}^2 \delta_k. \end{aligned}$$

Plugging (6.46) and (6.45) into (6.44) yields the desired statement. □

In the next lemma it remains to bound δ_k.

Lemma 6.4.7. *We have that for all $k \geq 0$, then*

$$\delta_{k+1} \leq \left(1 - \frac{1}{2\kappa}\right) \delta_k + \kappa^3 \|x_{k+1} - x_k\|^2.$$

Proof. Let $k \geq 0$ be fixed. From the definition of y_{k+1}, see (6.43), we deduce that

$$\begin{aligned} \delta_{k+1} &= \|y_{k+1}^* - y_{k+1}\|^2 \\ &= \|\operatorname{prox}_{\eta_y h} \left(y_{k+1}^* + \eta_y \nabla_y \Phi(x_{k+1}, y_{k+1}^*)\right) - \operatorname{prox}_{\eta_y h} \left(y_k + \eta_y \nabla_y \Phi(x_{k+1}, y_k)\right) \|^2. \end{aligned}$$

If Φ is strongly concave in its second component we can use the nonexpansiveness of the proximal operator and Lemma 2.2.3, which states

$$\begin{aligned} &\langle \nabla_y \Phi(x_{k+1}, y_{k+1}^*) - \nabla_y \Phi(x_{k+1}, y_k), y_{k+1}^* - y_k \rangle \\ &\quad \leq -\frac{\mu L_{\nabla \Phi}}{\mu + L_{\nabla \Phi}} \|y_{k+1}^* - y_k\|^2 - \frac{1}{\mu + L_{\nabla \Phi}} \|\nabla_y \Phi(x_{k+1}, y_{k+1}^*) - \nabla_y \Phi(x_{k+1}, y_k)\|^2 \end{aligned} \tag{6.47}$$

to conclude

$$\begin{aligned} \delta_{k+1} &\leq \|y_{k+1}^* + \eta_y \nabla_y \Phi(x_{k+1}, y_{k+1}^*) - y_k - \eta_y \nabla_y \Phi(x_{k+1}, y_k)\|^2 \\ &= \|y_{k+1}^* - y_k\|^2 + 2\eta_y \langle y_{k+1}^* - y_k, \nabla_y \Phi(x_{k+1}, y_{k+1}^*) - \nabla_y \Phi(x_{k+1}, y_k) \rangle \\ &\quad + \eta_y^2 \|\nabla_y \Phi(x_{k+1}, y_{k+1}^*) - \nabla_y \Phi(x_{k+1}, y_k)\|^2 \\ &\overset{(6.47)}{\leq} \left(\frac{\kappa - 1}{\kappa + 1}\right) \|y_{k+1}^* - y_k\|^2 \leq q \|y_{k+1}^* - y_k\|^2 \end{aligned}$$

with $q := \left(\frac{\kappa}{\kappa+1}\right)^2$, where we used that $\eta_y = 1/L_{\nabla \Phi}$. If on the other hand $-h$ is strongly concave we can use the fact that the proximal operator (of h) is even a contraction, see [6, Proposition 25.9 (i)], to deduce that

$$\delta_{k+1} \leq q \|y_{k+1}^* - y_k\|^2.$$

Therefore, in either case $\delta_{k+1} \leq q\|y^*_{k+1} - y_k\|^2$. Using this, the triangle inequality and Young's inequality, we have

$$\begin{aligned}
\delta_{k+1} &\leq q\|y^*_{k+1} - y_k\|^2 \\
&\leq q\Big(\|y^*_{k+1} - y^*_k\| + \|y^*_k - y_k\|\Big)^2 \\
&\leq q\left(1 + \frac{3\kappa^2 - 1}{2\kappa^3}\right) \underbrace{\|y^*_k - y_k\|^2}_{=\delta_k} + q\left(1 + \frac{2\kappa^3}{3\kappa^2 - 1}\right)\|y^*_{k+1} - y^*_k\|^2 \\
&\leq \left(1 - \frac{1}{2\kappa}\right)\delta_k + \kappa\|y^*_{k+1} - y^*_k\|^2. \qquad (6.48)
\end{aligned}$$

Due to the κ-Lipschitz continuity of $y^*(\cdot)$ we have that $\|y^*_{k+1} - y^*_k\| \leq \kappa\|x_{k+1} - x_k\|$, which finishes the proof. □

Now we can bound the sum of δ_k.

Lemma 6.4.8. *We have that, for all $K \geq 1$*

$$\sum_{k=0}^{K-1} \delta_k \leq 2\kappa\delta_0 + 2\kappa^4 \sum_{k=0}^{K-1} \|x_{k+1} - x_k\|^2.$$

Proof. By recursively applying the previous lemma we obtain for $k \geq 1$

$$\delta_k \leq \left(1 - \frac{1}{2\kappa}\right)^k \delta_0 + \kappa^3 \sum_{j=0}^{k-1} \left(1 - \kappa\frac{1}{2\kappa}\right)^{k-j-1} \|x_{j+1} - x_j\|^2.$$

Now we sum this inequality from $k = 1$ to $K - 1$ and add δ_0 on both sides to deduce that

$$\sum_{k=0}^{K-1} \delta_k \leq 2\kappa\delta_0 + 2\kappa^4 \sum_{k=0}^{K-1} \|x_{k+1} - x_k\|^2,$$

where we used that

$$\sum_{k=1}^{K-1}\sum_{j=0}^{k-1} \left(1 - \frac{1}{2\kappa}\right)^{k-1-j} \|x_{j+1} - x_j\|^2 \leq \left(\sum_{j=0}^{K-1} \left(1 - \frac{1}{2\kappa}\right)^j\right)\left(\sum_{k=0}^{K-1} \|x_{k+1} - x_k\|^2\right), \qquad (6.49)$$

and $\sum_{j=0}^{\infty} \left(1 - (2\kappa)^{-1}\right)^j = 2\kappa$. □

We can now put the pieces together.

Proof of Theorem 6.4.5. Summing up the inequality of Lemma 6.4.6 from $k = 0$ to $K - 1$ and applying Lemma 6.4.12 we deduce that

$$\begin{aligned}
\frac{1}{2}\eta_x \sum_{k=1}^{K} \|w_k\|^2 &\leq g(x_0) - g(x_K) \\
&\quad + \frac{1}{2}\Big(L_{\nabla\varphi} + 2L^2_{\nabla\varphi}\eta_x - \frac{1}{\eta_x} + 2\kappa^4 L^2_{\nabla\Phi}\eta_x\Big) \sum_{k=0}^{K-1} \|x_{k+1} - x_k\|^2 + 2L^2_{\nabla\Phi}\eta_x\kappa\delta_0.
\end{aligned}$$

With the stepsize $\eta_x = 1/(3(\kappa+1)^2 L_{\nabla\Phi})$ it follows that

$$
\begin{aligned}
L_{\nabla\varphi} + 2L_{\nabla\varphi}^2\eta_x - \frac{1}{\eta_x} + 2\kappa^4 L_{\nabla\Phi}^2\eta_x &= (\kappa+1)L_{\nabla\Phi} + \frac{2L_{\nabla\Phi}}{3} - 3(\kappa+1)^2 L_{\nabla\Phi} + \frac{2\kappa^4 L_{\nabla\Phi}}{3(\kappa+1)^2} \\
&\leq -\frac{2}{3}(\kappa+1)^2 L_{\nabla\Phi} \leq 0,
\end{aligned}
$$

which concludes the proof. □

6.4.3 Stochastic setting

For the purpose of this section Algorithm 6.2.3 reads as

$$
(\forall k \geq 0) \left\lfloor \begin{array}{l} x_{k+1} = \operatorname{prox}_{\eta_x f}\left(x_k - \eta_x G_x\right) \\ y_{k+1} = \operatorname{prox}_{\eta_y h}\left(y_k + \eta_y G_y\right), \end{array} \right. \tag{6.50}
$$

where we denote $G_x = \frac{1}{M}\sum_{i=1}^M \nabla_x \Phi(x_k, y_k; \xi_k^i)$ and $G_y = \frac{1}{M}\sum_{i=1}^M \nabla_y \Phi(x_{k+1}, y_k; \zeta_k^i)$.

Theorem 6.4.9. *Let in addition to the assumptions of Theorem 6.4.5 also the two properties of the gradient estimator Assumption 6.2.1 and 6.2.2 hold true. The iterates generated by* (6.50) *with stepsize $\eta_y = 1/L_{\nabla\Phi}$ and $\eta_x = 1/(4(1+\kappa)^2 L_{\nabla\Phi})$ and batch size $M = \mathcal{O}(\kappa\epsilon^{-2})$ fulfill*

$$
\min_{1\leq k\leq K} \mathbb{E}\left[\operatorname{dist}\Big(-\nabla\varphi(x_k), \partial f(x_k)\Big)^2\right] \leq 2\frac{\Delta^*}{\eta_x K} + 4\frac{L_{\nabla\Phi}^2\kappa\|y^*(x_0) - y_0\|^2}{K} + 4\epsilon^2(\kappa+1)\sigma^2,
$$

where $\Delta^* = g(x_0) - \inf_{x\in\mathbb{R}^d} g(x)$. This means that a ϵ-stationary point is visited in at most $\mathcal{O}(\kappa^2\epsilon^{-2})$ iterations resulting in $\mathcal{O}(\kappa^3\epsilon^{-4})$ stochastic gradient evaluations.

Let us start with the first lemma.

Lemma 6.4.10. *There exists a sequence of points $(w_k)_{k\geq 1}$ such that $w_k \in (\partial f + \nabla\varphi)(x_k)$ and its norm can be bounded by*

$$
\begin{aligned}
\frac{1}{2}\eta_x\mathbb{E}\big[\|w_{k+1}\|^2\big] \leq \mathbb{E}[g(x_k) - g(x_{k+1})] &+ \frac{1}{2}\Big(L_{\nabla\varphi} + 3L_{\nabla\varphi}^2\eta_x - \frac{1}{\eta_x}\Big)\mathbb{E}\big[\|x_k - x_{k+1}\|^2\big] \\
&+ \eta_x L_{\nabla\Phi}^2\mathbb{E}[\delta_k] + \eta_x\frac{\sigma^2}{M}
\end{aligned}
$$

for all $k \geq 0$.

Proof. Let $k \geq 0$. From the definition of the proximal operator we deduce that

$$
0 \in \partial f(x_{k+1}) + G_x + \frac{1}{\eta_x}(x_{k+1} - x_k).
$$

Thus,

$$
w_{k+1} := \frac{1}{\eta_x}(x_k - x_{k+1}) + \nabla\varphi(x_{k+1}) - G_x \in \partial f(x_{k+1}) + \nabla\varphi(x_{k+1}).
$$

In order to bound w_{k+1} we consider

$$\begin{aligned}&\|w_{k+1}\|^2\\ &\quad= \eta_x^{-2}\|x_k - x_{k+1}\|^2 + 2\eta_x^{-1}\langle x_k - x_{k+1}, \nabla\varphi(x_{k+1}) - G_x\rangle + \|\nabla\varphi(x_{k+1}) - G_x\|^2.\end{aligned} \tag{6.51}$$

Analogously to (6.46) we have that

$$\langle x_k - x_{k+1}, \nabla\varphi(x_{k+1}) - G_x\rangle \leq g(x_k) - g(x_{k+1}) + \frac{1}{2}\Big(L_{\nabla\varphi} - \frac{2}{\eta_x}\Big)\|x_{k+1} - x_k\|^2. \tag{6.52}$$

Using (6.52) and

$$\begin{aligned}&\mathbb{E}\big[\|\nabla\varphi(x_{k+1}) - G_x\|^2\big]\\ &\quad= \mathbb{E}\big[\|\nabla\varphi(x_{k+1}) - \nabla\varphi(x_k) + \nabla\varphi(x_k) - \nabla_x\Phi(x_k, y_k) + \nabla_x\Phi(x_k, y_k) - G_x\|^2\big]\\ &\quad= \mathbb{E}\big[\|\nabla\varphi(x_{k+1}) - \nabla\varphi(x_k)\|^2 + 2\langle\nabla\varphi(x_{k+1}) - \nabla\varphi(x_k), \nabla\varphi(x_k) - \nabla_x\Phi(x_k, y_k)\rangle\big]\\ &\qquad + 2\mathbb{E}[\langle\nabla\varphi(x_{k+1}) - \nabla\varphi(x_k), \nabla_x\Phi(x_k, y_k) - G_x\rangle]\\ &\qquad + 2\underbrace{\mathbb{E}[\langle\nabla\varphi(x_k) - \nabla_x\Phi(x_k, y_k), \nabla_x\Phi(x_k, y_k) - G_x\rangle]}_{=0}\\ &\qquad + \mathbb{E}\big[\|\nabla\varphi(x_k) - \nabla_x\Phi(x_k, y_k)\|^2\big] + \mathbb{E}\big[\|\nabla_x\Phi(x_k, y_k) - G_x\|^2\big]\\ &\quad\leq 3L_{\nabla\varphi}^2\mathbb{E}\big[\|x_{k+1} - x_k\|^2\big] + 2L_{\nabla\Phi}^2\mathbb{E}[\delta_k] + 2\frac{\sigma^2}{M}\end{aligned}$$

in (6.51) yields the desired statement. □

In the next lemma it remains to bound δ_k.

Lemma 6.4.11. *We have that for all $k \geq 0$*

$$\mathbb{E}[\delta_{k+1}] \leq \Big(1 - \frac{1}{2\kappa}\Big)\mathbb{E}[\delta_k] + \kappa^3\|x_{k+1} - x_k\|^2 + \frac{\sigma^2}{ML_{\nabla\Phi}^2}.$$

Proof. Let $k \geq 0$ be fixed. We first consider the case where Φ is strongly concave in its second component from the definition of y_{k+1} (see (6.50)) we deduce that

$$\begin{aligned}\delta_{k+1} &= \|y_{k+1}^* - y_{k+1}\|^2\\ &= \|\mathrm{prox}_{\eta_y h}\left(y_{k+1}^* + \eta_y\nabla_y\Phi(x_{k+1}, y_{k+1}^*)\right) - \mathrm{prox}_{\eta_y h}\left(y_k + \eta_y G_y\right)\|^2\\ &\leq \|y_{k+1}^* + \eta_y\nabla_y\Phi(x_{k+1}, y_{k+1}^*) - y_k - \eta_y G_y\|^2\end{aligned}$$

and

$$\begin{aligned}&\|y_{k+1}^* + \eta_y\nabla_y\Phi(x_{k+1}, y_{k+1}^*) - y_k - \eta_y G_y\|^2\\ &\quad= \|y_{k+1}^* - y_k\|^2 + 2\eta_y\langle y_{k+1}^* - y_k, \nabla_y\Phi(x_{k+1}, y_{k+1}^*) - \nabla_y\Phi(x_{k+1}, y_k)\rangle\\ &\qquad + 2\eta_y\underbrace{\langle y_{k+1}^* - y_k, \nabla_y\Phi(x_{k+1}, y_k) - G_y\rangle}_{(\square)}\\ &\qquad + \eta_y^2\|\nabla_y\Phi(x_{k+1}, y_{k+1}^*) - \nabla_y\Phi(x_{k+1}, y_k)\|^2 + \eta_y^2\|\nabla_y\Phi(x_{k+1}, y_k) - G_y\|^2\\ &\qquad + \eta_y^2\underbrace{\langle\nabla_y\Phi(x_{k+1}, y_{k+1}^*) - \nabla_y\Phi(x_{k+1}, y_k), \nabla_y\Phi(x_{k+1}, y_k) - G_y\rangle}_{(*)}.\end{aligned} \tag{6.53}$$

Some of the terms vanish after taking the expectation such as $\mathbb{E}[(*)] = \mathbb{E}[\mathbb{E}[(*) \,|\, y_k, x_{k+1}]] = \mathbb{E}[0] = 0$ and $\mathbb{E}[(\square)] = \mathbb{E}[\mathbb{E}[(\square) \,|\, y_k, x_{k+1}]] = 0$. Using furthermore Lemma 2.2.3 which states that

$$\begin{aligned}&\langle \nabla_y \Phi(x_{k+1}, y^*_{k+1}) - \nabla_y \Phi(x_{k+1}, y_k), y^*_{k+1} - y_k \rangle \\ &\qquad \leq -\frac{\mu L_{\nabla\Phi}}{\mu + L_{\nabla\Phi}} \|y^*_{k+1} - y_k\|^2 - \frac{1}{\mu + L_{\nabla\Phi}} \|\nabla_y \Phi(x_{k+1}, y^*_{k+1}) - \nabla_y \Phi(x_{k+1}, y_k)\|^2\end{aligned} \tag{6.54}$$

results in

$$\mathbb{E}[\delta_{k+1}] \leq \left(\frac{\kappa - 1}{\kappa + 1}\right) \mathbb{E}\big[\|y^*_{k+1} - y_k\|^2\big] + \frac{\sigma^2}{M L^2_{\nabla\Phi}} \leq q\mathbb{E}\big[\|y^*_{k+1} - y_k\|^2\big] + \frac{\sigma^2}{M L^2_{\nabla\Phi}}, \tag{6.55}$$

with $q = \left(\frac{\kappa}{\kappa+1}\right)^2$, where we used that $\eta_y = {}^1\!/\!{}_{L_{\nabla\Phi}}$. If h is strongly concave then we use the fact that the proximal operator is a contraction, see [6, Proposition 23.11], to deduce that

$$\begin{aligned}\mathbb{E}[\delta_{k+1}] &= \mathbb{E}\big[\|y^*_{k+1} - y_{k+1}\|^2\big] \\ &= \mathbb{E}\Big[\|\mathrm{prox}_{\eta_y h}\left(y^*_{k+1} + \eta_y \nabla_y \Phi(x_{k+1}, y^*_{k+1})\right) - \mathrm{prox}_{\eta_y h}\left(y_k + \eta_y G_y\right)\|^2\Big] \\ &= q\mathbb{E}\big[\|y^*_{k+1} - y_k + \eta_y \nabla_y \Phi(x_{k+1}, y^*_{k+1}) - \eta_y G_y\|^2\big] \\ &\overset{(6.53)}{=} q\mathbb{E}\big[\|y^*_{k+1} - y_k\|^2\big] + 2q\eta_y \mathbb{E}\big[\langle y^*_{k+1} - y_k, \nabla_y \Phi(x_{k+1}, y^*_{k+1}) - \nabla_y \Phi(x_{k+1}, y_k)\rangle\big] \\ &\quad + q\eta_y^2 \mathbb{E}\big[\|\nabla_y \Phi(x_{k+1}, y^*_{k+1}) - \nabla_y \Phi(x_{k+1}, y_k)\|^2 + q\eta_y^2 \|\nabla_y \Phi(x_{k+1}, y_k) - G_y\|^2\big]\end{aligned}$$

Using now (6.54) with $\mu = 0$, i.e. the cocoercivity of the gradient, we deduce that

$$\mathbb{E}[\delta_{k+1}] = \mathbb{E}\big[\|y^*_{k+1} - y_{k+1}\|^2\big] \leq q\|y^*_{k+1} - y_k\|^2 + q\frac{\sigma^2}{M L^2_{\nabla\Phi}},$$

meaning that we concluded (6.55) in both cases. Next, using (6.55) and the considerations made in (6.48) we deduce that

$$\mathbb{E}[\delta_{k+1}] \leq \left(1 - \frac{1}{2\kappa}\right) \mathbb{E}[\delta_k] + \kappa \mathbb{E}\big[\|y^*_{k+1} - y^*_k\|^2\big] + \frac{\sigma^2}{M L^2_{\nabla\Phi}}.$$

Again, due to the κ-Lipschitz continuity of $y^*(\cdot)$ we have that $\|y^*_{k+1} - y^*_k\| \leq \kappa \|x_{k+1} - x_k\|$, which finishes the proof. □

Now we can bound the sum of δ_k.

Lemma 6.4.12. *We have that, for all $K \geq 1$*

$$\sum_{k=0}^{K-1} \mathbb{E}[\delta_k] \leq 2\kappa\delta_0 + 2\kappa^4 \sum_{k=0}^{K-1} \mathbb{E}\big[\|x_{k+1} - x_k\|^2\big] + 2K \frac{\kappa\sigma^2}{M L^2_{\nabla\Phi}}.$$

Proof. By recursively applying the previous lemma we obtain for $k \geq 1$

$$\mathbb{E}[\delta_k] \leq \left(1 - \frac{1}{2\kappa}\right)^k \delta_0 + \sum_{j=0}^{k-1} \left(1 - \frac{1}{2\kappa}\right)^{k-j-1} \left(\kappa^3 \mathbb{E}\big[\|x_{j+1} - x_j\|^2\big] + \frac{\sigma^2}{M L_{\nabla\Phi}^2}\right).$$

Now we sum this inequality from $k = 1$ to $K - 1$ and add δ_0 on both sides to deduce that

$$\sum_{k=0}^{K-1} \mathbb{E}[\delta_k] \leq 2\kappa\delta_0 + 2K \frac{\kappa\sigma^2}{M L_{\nabla\Phi}^2} + 2\kappa^4 \sum_{k=0}^{K-1} \mathbb{E}\big[\|x_{k+1} - x_k\|^2\big]$$

using the considerations made in (6.49). □

We can now put the pieces together.

Proof of Theorem 6.4.9. We sum up the inequality of Lemma 6.4.10 from $k = 0$ to $K - 1$ and applying Lemma 6.4.12 we deduce that

$$\begin{aligned} \frac{1}{2}\eta_x \sum_{k=1}^{K} \mathbb{E}\big[\|w_k\|^2\big] \leq \mathbb{E}[g(x_0) - g(x_K)] + 2\eta_x\kappa L_{\nabla\Phi}^2 \delta_0 + 2\eta_x\kappa K \frac{\sigma^2}{M} + \eta_x K \frac{\sigma^2}{M} \\ + \frac{1}{2}\Big(L_{\nabla\varphi} + 3L_{\nabla\varphi}^2 \eta_x - \frac{1}{\eta_x} + 2\kappa^4 L_{\nabla\Phi}^2 \eta_x\Big) \sum_{k=0}^{K-1} \mathbb{E}\big[\|x_{k+1} - x_k\|^2\big]. \end{aligned}$$

Applying the stepsize $\eta_x = 1/(3(1+\kappa)^2 L_{\nabla\Phi})$ it follows that

$$\begin{aligned} L_{\nabla\varphi} + 3L_{\nabla\varphi}^2 \eta_x - \frac{1}{\eta_x} + 2\kappa^4 L_{\nabla\Phi}^2 \eta_x &\leq 2(\kappa+1)L_{\nabla\Phi} - 3(\kappa+1)^2 L_{\nabla\Phi} + \frac{2\kappa^2 L_{\nabla\Phi}}{3} \\ &\leq -\frac{1}{3}(\kappa+1)^2 L_{\nabla\Phi} \leq 0 \end{aligned}$$

which concludes the proof. □

7 Conclusion

In this thesis we investigated different structured nonsmooth optimization problems and iterative schemes to solve them. We studied the convergence behavior of these solution methods in terms of quantitative global bounds, i.e. convergence rates. This kind of worst case analysis is fundamental in giving a principled understanding of the performance of different algorithms. Where we deemed it appropriate, these theoretical considerations were augmented with numerical experiments.

In Chapter 3 and 4 we studied composite optimization problems where a nonsmooth function is composed with a linear operator and highlighted applications in imaging and machine learning. We focused on full splitting methods where the proximal operator of the outer nonsmooth function is evaluated separately from the matrix, resulting in easy to use algorithms.

In Chapter 3 we presented a novel (randomized) method for the convex formulation of the aforementioned problem using stochastic accelerated gradient evaluations of the Moreau envelope and proved state-of-the-art convergence guarantees. While the empirical performance of our proposed algorithm was mostly comparable to the primal-dual methods [27,28], it outperformed them significantly when used as a subroutine inside the prox-linear algorithm [37] for weakly convex problems.

Chapter 4 dealt with the same problem formulation but dropped the convexity assumption. This enables the use of more sophisticated regularizers [38,108], possibly reducing the bias caused by convex functions such as the 1-norm. While similar formulation have been considered before we proposed a novel method and proved a worst case complexity which interpolates nicely between gradient descent for smooth objectives and black-box subgradient descent for nonsmooth problems.

In Chapter 5 we highlighted the connection between GAN objectives and monotone inclusions and were therefore able to tackle their training via the Forward-Backward-Forward (FBF) method which is known to converge to a solution for convex-concave minimax problems. We deepened this theoretical understanding by proving novel convergence rates in terms of the function values. Since FBF provides a natural way to deal with nonsmooth regularizers via the proximal mapping, we modified the WGAN objective to encompass a 1-norm instead of the usual weight clipping. We showed that this formulation provides a benefit for all considered methods, smoothing the training process and improving Inception Score and Fréchet Inception Distance. Moreover, FBF outperformed all competitors including the commonly used gradient descent ascent (GDA) method as well as Extragradient [40] and Optimistic GDA [33].

Lastly, Chapter 6 was devoted to a theoretical study of the already well established GDA method. In particular, we considered its alternating variant, where the two components are updated in a sequential manner. In the convex setting is known that for equal stepsizes simultaneous GDA diverges while the alternating update scheme at least provides bounded trajectories. Although the assumptions in this chapter are sometimes rather restrictive, our obtained convergence rates were still novel outside the convex-concave setting, albeit only a slight improvement over simultaneous GDA was obtained [60].

Bibliography

[1] J. Adler, H. Kohr, and O. Öktem. Operator discretization library (ODL). *Software available from https://github. com/odlgroup/odl*, 2017.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv:1701.07875*, 2017.

[3] F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[4] J. P. Bailey, G. Gidel, and G. Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Proceedings of the 33rd Conference on Learning Theory*, pages 391–407. PMLR, 2020.

[5] S. Barratt and R. Sharma. A note on the inception score. *arXiv:1801.01973*, 2018.

[6] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

[7] I. Bayram. On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Transactions on Signal Processing*, 64(6):1597–1608, 2015.

[8] A. E. Beaton and J. W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[9] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[10] Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. In *Advances in Neural Information Processing Systems*, pages 123–130, 2006.

[11] A. Böhm and A. Daniilidis. Ubiquitous algorithms in convex optimization generate self-contracted sequences. *arXiv:2003.04201*, 2020.

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

[13] A. Böhm and S. J. Wright. Variable smoothing for weakly convex composite functions. *arXiv:2003.07612*, 2020.

[14] R. I. Boţ and A. Böhm. An incremental mirror descent subgradient algorithm with random sweeping and proximal step. *Optimization*, 68(1):33–50, 2019.

[15] R. I. Boţ and A. Böhm. Variable smoothing for convex optimization problems using stochastic gradients. *arXiv:1905.06553*, 2019.

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

[17] R. I. Boţ and E. R. Csetnek. On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems. *Optimization*, 64(1):5–23, 2015.

[18] R. I. Boţ, E. R. Csetnek, A. Heinrich, and C. Hendrich. On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. *Mathematical Programming*, 150(2):251–279, 2015.

[19] R. I. Boţ and C. Hendrich. A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems. *Computational Optimization and Applications*, 54(2):239–262, 2013.

[20] R. I. Boţ and C. Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization*, 23(4):2541–2565, 2013.

[21] R. I. Boţ and C. Hendrich. Convergence analysis for a primal-dual monotone+ skew splitting algorithm with applications to total variation minimization. *Journal of Mathematical Imaging and Vision*, 2014.

[22] R. I. Boţ and C. Hendrich. On the acceleration of the double smoothing technique for unconstrained convex optimization problems. *Optimization*, 64(2):265–288, 2015.

[23] R. I. Boţ and C. Hendrich. A variable smoothing algorithm for solving convex optimization problems. *TOP*, 23(1):124–150, 2015.

[24] R. I. Boţ, M. Sedlmayer, and P. T. Vuong. A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANs. *arXiv:2003.07886*, 2020.

[25] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004.

[26] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization theory and Applications*, 166(3):968–982, 2015.

[27] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 2018.

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

[29] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, pages 391–401, 2019.

[30] C. Chen, T. K. Pong, L. Tan, and L. Zeng. A difference-of-convex approach for split feasibility with applications to matrix factorizations and outlier detection. *Journal of Global Optimization*, pages 1–30, 2020.

[31] R. S. Chen, B. Lucier, Y. Singer, and V. Syrgkanis. Robust optimization for nonconvex objectives. In *Advances in Neural Information Processing Systems*, pages 4705–4714, 2017.

[32] L. Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 2013.

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

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

[35] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

[36] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[37] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

[38] J. Fan. Comments on «wavelets in statistics: A review» by a. antoniadis. *Journal of the Italian Statistical Society*, 6(2):131, 1997.

[39] Y. Fan, S. Lyu, Y. Ying, and B. Hu. Learning with average top-k loss. In *Advances in Neural Information Processing Systems*, pages 497–505, 2017.

[40] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[41] G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019.

[42] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv:1701.00160*, 2016.

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

[44] P. Groetzner and M. Dür. A factorization method for completely positive matrices. *Linear Algebra and its Applications*, 591:1–24, 2020.

[45] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[46] E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv:1803.01401*, 2018.

[47] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[49] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.

[50] T. Hoheisel, M. Laborde, and A. Oberman. On proximal point-type algorithms for weakly convex functions and their connection to the backward Euler method, 2018.

[51] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[52] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[53] W. Kong and R. D. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *arXiv:1905.13433*, 2019.

[54] G. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[55] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto, Canada*, 2009.

[56] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.

[57] G. Li and T. K. Pong. Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

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

[59] Q. Lin, M. Liu, H. Rafique, and T. Yang. Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities. *arXiv:1810.10207*, 2018.

[60] T. Lin, C. Jin, and M. I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv:1906.00331*, 2019.

[61] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. *arXiv:2002.02417*, 2020.

[62] M. Liu, Y. Mroueh, W. Zhang, X. Cui, T. Yang, and P. Das. Decentralized parallel algorithm for training generative adversarial nets. *arXiv:1910.12999*, 2019.

[63] P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(3):866–896, 2017.

[64] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 2020.

[65] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

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

[67] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019.

[68] L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, 2018.

[69] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330. Springer Science & Business Media, 2006.

[70] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2016.

[71] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[72] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[73] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[74] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[75] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[76] Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110(2):245–259, 2007.

[77] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[78] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14934–14942, 2019.

[79] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. *arXiv:1703.06182*, 2017.

[80] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *arXiv:2002.07919*, 2020.

[81] D. P. Palomar and Y. C. Eldar. *Convex Optimization in Signal Processing and Communications*. Cambridge university press, 2010.

[82] A. Parekh and I. W. Selesnick. Convex denoising using non-convex tight frame regularization. *IEEE Signal Processing Letters*, 22(10):1786–1790, 2015.

[83] J.-C. Pesquet and A. Repetti. A class of randomized primal-dual algorithms for distributed optimization. *arXiv:1406.6404*, 2014.

[84] K. Pieper and A. Petrosyan. Nonconvex penalization for sparse neural networks. *arXiv:2004.11515*, 2020.

[85] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

[86] H. Rafique, M. Liu, Q. Lin, and T. Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv:1810.02060*, 2018.

[87] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 993–1019, 2013.

[88] S. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[89] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[90] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

[91] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[92] L. Rosasco, S. Villa, and B. C. Vũ. A first-order stochastic primal-dual algorithm with correction step. *Numerical Functional Analysis and Optimization*, 38(5):602–626, 2017.

[93] S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu. ℓ 1 regularization in infinite dimensional feature spaces. In *International Conference on Computational Learning Theory*, pages 544–558. Springer, 2007.

[94] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[95] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[96] X. Shen and Y. Gu. Nonconvex sparse logistic regression with weakly convex regularization. *IEEE Transactions on Signal Processing*, 66(12):3199–3211, 2018.

[97] S. Shreve. *Stochastic Calculus for Finance I: The Binomial Asset Pricing Model.* Springer Science & Business Media, 2005.

[98] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[99] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pages 12680–12691, 2019.

[100] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[101] Q. Tran-Dinh, O. Fercoq, and V. Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28(1):96–134, 2018.

[102] Q. Tran-Dinh, D. Liu, and L. M. Nguyen. Hybrid variance-reduced SGD algorithms for nonconvex-concave minimax problems. *arXiv:2006.15266*, 2020.

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

[104] J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior (Commemorative Edition).* Princeton University Press, 2007.

[105] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 2013.

[106] Z. Xu, H. Zhang, Y. Xu, and G. Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv:2006.02032*, 2020.

[107] Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pages 451–459, 2016.

[108] C.-H. Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[109] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363. PMLR, 2019.

[110] R. Zhao. A primal dual smoothing framework for max-structured nonconvex optimization. *arXiv:2003.04375*, 2020.