An Adaptive Mirror-Prox Algorithm for Variational Inequalities with Singular Operators

Kimon Antonakopoulos
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP
LIG 38000 Grenoble, France.
kimon.antonakopoulos@inria.fr

E. Veronica Belmega
ETIS/ENSEA
Univ. de Cergy-Pontoise-CNRS, France
belmega@ensea.fr

Panayotis Mertikopoulos
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP,
LIG 38000 Grenoble, France.
panayotis.mertikopoulos@imag.fr

Abstract

Lipschitz continuity is a central requirement for achieving the optimal $\mathcal{O}(1/T)$ rate of convergence in monotone, deterministic variational inequalities (a setting that includes convex minimization, convex-concave optimization, nonatomic games, and many other problems). However, in many cases of practical interest, the operator defining the variational inequality may exhibit singularities at the boundary of the feasible region, precluding in this way the use of fast gradient methods that attain this optimal rate (such as Nemirovski's mirror-prox algorithm and its variants). To address this issue, we propose a novel regularity condition which we call *Bregman continuity*, and which relates the variation of the operator to that of a suitably chosen Bregman function. Leveraging this condition, we derive an adaptive mirror-prox algorithm which attains the optimal $\mathcal{O}(1/T)$ rate of convergence in problems with possibly singular operators, without any prior knowledge of the degree of smoothness (the Bregman analogue of the Lipschitz constant). We also show that, under Bregman continuity, the mirror-prox algorithm achieves a $\mathcal{O}(1/\sqrt{T})$ convergence rate in stochastic variational inequalities.

1 Introduction

The seminal introduction of generative adversarial networks (GANs) [17] has ushered in a new optimization paradigm in deep learning: instead of focusing on the minimization of an empirical loss function, GAN training hinges on a zero-sum game between a generator and a discriminator. In fact, in many cases GAN training goes even beyond the min-max setting, either because there are more than two networks involved, or because the objectives of the generator and the discriminator are not entirely opposed – e.g., as in the widely used ACGAN framework of Odena et al. [41]. In these cases, the most compact way of representing the problem's training landscape is by means of a *variational inequality* (VI).

Tracing their origins to the work of Stampacchia [47] on the Signorini problem, variational inequalities have since found a broad range of applications in physics, engineering, economics – and, more recently, machine learning. One of the main reasons for their extensive applicability is that they comprise a flexible optimization framework which can simultaneously account for loss function minimization, saddle-point, game-theoretic, and fixed point problems. As a result, there has been considerable interest in the literature to develop optimal algorithms for solving VI problems; for an appetizing introduction, see [14] and references therein.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

One of the most widely studied methods for this purpose is ordinary gradient descent – also known as the forward-backward (FB) algorithm in operator theory [6].[1] In monotone, deterministic variational inequalities, the convergence of the method is guaranteed under a condition known as *cocoercivity*. By the Baillon–Haddad theorem, if the operator defining the variational inequality is a gradient field (i.e., in loss minimization problems), this condition is equivalent to Lipschitz smoothness of the associated loss function [4, 6]. However, cocoercivity may fail to hold even in simple, bilinear min-max problems, in which case gradient descent provably fails to converge – see e.g., [16, 33, 34] for a precise statement.

The first algorithm achieving convergence in (pseudo-)monotone variational inequalities without cocoercivity is the *extra-gradient* (EG) algorithm of Korpelevich [23], which only requires Lipschitz continuity of the underlying operator.[2] The asymptotic convergence result of Korpelevich [23] was subsequently extended by Nemirovski [36] who introduced the *mirror-prox* (MP) algorithm, a Bregman variant of the EG algorithm with ergodic averaging. As was shown in [36], the mirror-prox algorithm attains a $\mathcal{O}(1/T)$ ergodic convergence rate in monotone variational inequalities with Lipschitz continuous operators, and this rate cannot be improved without further assumptions.

However, in many applications and problems of practical interest, Lipschitz continuity may also fail to hold, either because the loss profile of the problem grows too rapidly (e.g., as in support vector machines or GAN models with Kullback-Leibler losses), or because the problem exhibits singularities near the boundary of the feasible region (e.g., as in resource allocation and inverse problems). In these cases, one would still want to apply a fast method like mirror-prox, but the lack of smoothness means that there are no convergence guarantees – asymptotically, ergodically, or otherwise.

Our contributions. Our starting point is the observation that this failure stems from the fact that Lipschitz continuity of the operator is defined relative to a *global* norm. Because of this, the standard Lipschitz framework is not well-suited to problems with singularities or rapid growth: a global norm is oblivious to the geometry of the feasible region (and, in particular, its boundary), so it cannot capture the finer features of the problem's loss landscape.

To overcome this limitation, we introduce a novel regularity condition, which we call *Bregman continuity*, and which is made-to-order for the singularity landscape of the problem at hand. Specifically, instead of defining Lipschitz continuity relative to a global norm, we define it in terms of a family of *local* norms and a suitably chosen *Bregman function*. This leads to an intricate interplay between different geometric notions of distance (the Bregman divergence and the local norm), but it also introduces the flexibility required to tackle variational inequalities with singular operators.

Under this assumption, we show that the mirror-prox algorithm attains the optimal $\mathcal{O}(1/T)$ convergence rate in variational inequalities with (possibly) singular operators, provided that the method is run with the same Bregman function that is used to define Bregman continuity. As in the standard Lipschitz framework, the method's convergence requires a step-size of the form $\gamma < 1/\beta$, where β is the Bregman constant of the operator (i.e., the Bregman analogue of the Lipschitz constant). Estimating this constant can be fairly challenging in practice (if not downright impossible), so we also introduce an *adaptive mirror-prox* (AMP) method which attains the same $\mathcal{O}(1/T)$ rate without requiring any a priori estimation of β – essentially, the Bregman constant is learned at the same time as the problem's landscape. Finally, we provide a variant of the method for stochastic variational inequalities, and we establish a $\mathcal{O}(1/\sqrt{T})$ convergence rate in this setting. To the best of our knowledge, these are the first results of this kind in the literature.

Related work. Owing to their optimal rate guarantees, the extra-gradient and mirror-prox algorithms have been at the forefront of an extensive literature which is impossible to adequately review here. As a purely indicative list of contributions in the Lipschitz continuous setting (and with no illusion of being comprehensive), we refer the reader to Juditsky et al. [20], Chambolle and Pock [10], Malitsky [27], Iusem et al. [19] and Mokhtari et al. [35] for some recent developments. Especially in

[1] When used to find a zero of a composite operator, the FB algorithm is known as a "splitting" method; see e.g., Bruck Jr. [9], Passty [42], [12], and references therein.

[2] An operator $A(x)$ is cocoercive if $\langle A(x') - A(x), x' - x\rangle \geq (1/\beta)\|A(x') - A(x)\|^2$ for some $\beta > 0$ and all x, x'. Note that Lipschitz continuity is strictly weaker than cocoercivity: the operator $A(x_1, x_2) = (-x_2, x_1)$ is Lipschitz continuous over $\mathbb{R}^2$, but it is not cocoercive; see Section 2 for a detailed discussion.

learning theory, there has been a surge of interest motivated by the application of EG/MP methods to GAN training, see e.g., [13, 15, 34, 49] and references therein.[3]

Going beyond the Lipschitz regime, Bauschke et al. [7] recently introduced a "Lipschitz-like" smoothness condition for convex minimization problems and used it to establish a $\mathcal{O}(1/T)$ value convergence rate for mirror descent methods (as opposed to mirror-prox). Always in the context of loss minimization problems, Bolte et al. [8] subsequently extended the results of Bauschke et al. [7] to unconstrained non-convex problems that satisfy the Kurdyka–Łojasiewicz (KL) inequality, while Lu et al. [26] considered functions that are also relatively strongly convex and showed that mirror descent achieves a geometric convergence rate in this context. Finally, in a very recent preprint, Hanzely et al. [18] examined the rate of convergence of an accelerated variant of mirror descent under the same Lipschitz-like smoothness assumption.

The condition of Bauschke et al. [7] is remarkably simple as it only posits that the problem's loss function f is such that $\beta h - f$ is convex for some reference Bregman function h and some $\beta > 0$. A straightforward extension of this condition to an operator setting would be to require the monotonicity of $\beta\nabla h - A$, where A is the operator defining the variational inequality under study. However, the cornerstone of this "Lipschitz-like" condition is a descent lemma which does not carry over to variational inequalities, so it does not seem possible to extend the analysis of Bauschke et al. [7] to an operator setting. Lu [25] also considered a "relative continuity" condition for loss minimization problems positing that $\|\nabla f(x)\| \leq M \inf_{x'} \sqrt{2D(x',x)}/\|x' - x\|$ (where f is the problem's objective and D is the Bregman divergence of h). Written this way, the condition of Lu [25] can also be extended to an operator setting, but this would provide a surrogate for operator *boundedness*, not Lipschitz continuity (since $A = \nabla f$ in minimization problems). Since the optimal $\mathcal{O}(1/T)$ convergence rate of the mirror-prox algorithm is tied to the *regularity* of A – as opposed to its boundedness – the condition of Lu [25] does not seem applicable to the setting under study. Accordingly, there is no overlap in results or methodology with this particular strand of the literature.

Finally, in a very recent paper, Bach and Levy [3] introduced a *universal* variant of the mirror-prox algorithm which is model-agnostic and achieves an optimal convergence rate in stochastic and/or smooth settings. Achieving optimal rates in the setting of Bach and Levy [3] relies crucially on the operator being Lipschitz continuous (albeit with a possibly unknown constant) and the feasible region having a finite Bregman diameter. The algorithm we propose in this work is not universal but it *is* adaptive, and it does not require either Lipschitz continuity or a finite Bregman diameter. In this manner, our work also provides an important first step towards extending the universal analysis of Bach and Levy [3] to VI problems with singularities.

2 Preliminaries

Let $\mathcal{X}$ be a convex – but not necessarily closed or compact – subset of a d-dimensional normed space $\mathcal{V}$, and let $\mathcal{V}^*$ denote the dual space of $\mathcal{V}$. The *variational inequality* (VI) problem associated to a continuous operator $A\colon \mathcal{X} \to \mathcal{V}^*$ consists of finding $x^* \in \mathcal{X}$ such that

$$\langle A(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{VI}$$

Following [14], we will refer to this problem as $\mathrm{VI}(\mathcal{X}, A)$ and we will write $\mathcal{X}^* \equiv \mathrm{Sol}(\mathcal{X}, A)$ for its set of solutions. Note also that, if $\mathcal{X}$ is not closed, A may exhibit a *singularity* at a residual point $x \in \mathrm{bd}(\mathcal{X}) \setminus \mathcal{X}$ in the sense that A does not admit a continuous extension to x.

In the literature, this formulation of the problem is often reffered to as a *Stampacchia variational inequality* (SVI) [14] or a "strong" variational inequality [20, 38]. For illustration purposes, we present some archetypal examples of such problems below:

Example 2.1 (Loss minimization)**.** If $A = \nabla f$ for some convex loss function f on $\mathcal{X} = \mathbb{R}^d$, solutions of (VI) coincide with the global minimizers of f.

Example 2.2 (Min-max optimization)**.** Suppose that $A = (\nabla_{x_1} f, -\nabla_{x_2} f)$ for some real-valued function $f(x_1, x_2)$ with $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$. If f is convex-concave (i.e., convex in x_1 and concave in

[3] We note here that the method is sometimes referred to as "optimistic mirror descent". This terminology is due to Rakhlin and Sridharan [43, 44] and may refer either to the mirror-prox method itself, or to a variant with "gradient extrapolation from the past", as in [15].

x_2), any solution $x^* = (x_1^*, x_2^*)$ of (VI) is a global saddle-point of f, i.e.,

$$f(x_1^*, x_2^*) \leq f(x_1, x_2^*) \quad \text{and} \quad f(x_1^*, x_2^*) \geq f(x_1^*, x_2) \tag{2.1}$$

for all $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$. Problems of this type have attracted considerable interest in the fields of machine learning and artificial intelligence because they constitute the basic optimization framework for GANs [17]. For a series of recent papers focusing on the interplay between GAN and saddle-point problems / variational inequalities, see [13, 15, 24, 34, 49] and references therein.

Example 2.3 (Resource sharing problems)**.** Consider a set of *resources* $r \in \mathcal{R} = \{1, \dots, R\}$ serving a stream of *demands* that arrive at a rate of ρ per unit of time (for instance, a GPU cluster or a computing grid processing a stream of jobs). If the load on the r-th resource is x_r, the expected service time in the standard Kleinrock model [22] is given by the M/M/1 loss function

$$\ell_r(x_r) = \frac{1}{c_r - x_r}, \tag{2.2}$$

where c_r denotes the capacity of the resource. In this setting, the set of feasible resource allocations is $\mathcal{X} \equiv \{(x_1, \dots, x_R) : 0 \leq x_r < c_r, x_1 + \cdots + x_R = \rho\}$,[4] and we say that a resource allocation profile $x^* \in \mathcal{X}^*$ is at *Nash/Wardrop equilibrium* [40, 48] if

$$\ell_r(x_r^*) \leq \ell_r(x_r) \quad \text{for all } x \in \mathcal{X} \text{ and all } r \in \mathcal{R} \text{ such that } x_r^* > 0 \tag{2.3}$$

i.e., when no job would be better served by transferring it to a different priority queue. In this case, if we let $A(x) = (\ell_1(x_1), \dots, \ell_R(x_R))$, a standard calculation shows that x^* is an equilibrium allocation if and only if it solves the associated variational inequality problem for A.

The most widely used assumption in the literature for solving VI problems is *monotonicity*, i.e.,

$$\langle A(x') - A(x), x' - x \rangle \geq 0 \quad \text{for all } x, x' \in \mathcal{X}. \tag{2.4}$$

When $A = \nabla f$, this condition is equivalent to f being convex; likewise, when $A = (\nabla_{x_1} f, -\nabla_{x_2} f)$ as in Example 2.2, monotonicity is equivalent to f being convex-concave [6]; finally, by direct calculation, it is straightforward to see that the operator defined in Example 2.3 is monotone. For an introduction to the theory of monotone operators, we refer the reader to Facchinei and Pang [14] and Bauschke and Combettes [6].

Now, drawing on Nesterov [38, 39] and Juditsky et al. [20], if A is monotone, the quality of a candidate solution $\hat{x} \in \mathcal{X}$ can be assessed via the *restricted gap* (or *merit*) *function*

$$\mathrm{Gap}_{\mathcal{C}}(\hat{x}) = \sup_{x \in \mathcal{C}} \langle A(x), \hat{x} - x \rangle, \tag{2.5a}$$

where $\mathcal{C}$ is a nonempty convex subset of $\mathcal{X}$. The rationale behind this definition is that, if x^* solves (VI), monotonicity gives $\langle A(x), x^* - x \rangle \leq \langle A(x^*), x^* - x \rangle \leq 0$, so the quantity being maximized in (2.5) is small if $\hat{x}$ is an approximate solution of (VI). Formally, we have:

Lemma 1. *Suppose that A is monotone. If x^* solves* (VI)*, we have $\mathrm{Gap}_{\mathcal{C}}(x^*) = 0$ whenever $x^* \in \mathcal{C}$. Conversely, if $\mathrm{Gap}_{\mathcal{C}}(\hat{x}) = 0$ and $\mathcal{C}$ contains a neighborhood of $\hat{x}$ in $\mathcal{X}$, $\hat{x}$ is a solution of* (VI)*.*

This lemma extends a similar result by Nesterov [38], so we defer its proof to the paper's supplement. In view of all this, we will employ the gap function $\mathrm{Gap}_{\mathcal{C}}(\hat{x})$ as our main figure of merit and we will use it to state our convergence rate guarantees in the sequel.

3 Bregman continuity

In addition to monotonicity, a standard assumption for solving variational inequalities is *Lipschitz continuity*, i.e.,

$$\|A(x') - A(x)\|_* \leq \beta \|x' - x\| \tag{Lip}$$

for some $\beta > 0$ and for all $x, x' \in \mathcal{X}$. This definition involves two distinct (but related) measures of distance: (*i*) the primal norm on $\mathcal{V}$ which measures distances between the primal points $x, x' \in \mathcal{X}$; and (*ii*) the dual norm on $\mathcal{V}^*$ which measures the distance between the dual vectors $A(x), A(x') \in \mathcal{V}^*$.[5]

Importantly, both of these notions are *global*, i.e., they do not depend on the point in space at which they are calculated; as such, Lipschitz continuity is oblivious to the geometry of $\mathcal{X}$ (and, in particular, its boundary). In the sequel, we describe a way to overcome this limitation by introducing two distinct notions of distance that are tailored to the geometry of $\mathcal{X}$ and the singularity landscape of A.

[4] For posterity, note here that $\mathcal{X}$ is convex but it is not necessarily closed.

[5] Recall here that the dual norm of $v \in \mathcal{V}^*$ is defined as $\|v\|_* = \max_{z \in \mathcal{V}} \{|\langle v, z \rangle| : \|z\| \leq 1\}$.

Local norms. The first measure of distance that we define is that of *local norm* on $\mathcal{X}$:

Definition 1. Let $\mathcal{Z} = \operatorname{span}(\mathcal{X} - \mathcal{X})$ denote the *tangent hull* of $\mathcal{X}$, i.e., the subspace of $\mathcal{V}$ spanned by all possible displacement vectors of the form $z = x' - x$, $x, x' \in \mathcal{X}$. A *local norm* on $\mathcal{X}$ is a continuous assignment of a norm $\|\cdot\|_x$ on $\mathcal{Z}$ at each $x \in \mathcal{X}$.[6] The induced *dual local norm* is then defined as

$$\|v\|_{x,*} = \max_{z \in \mathcal{Z}} \{|\langle v, z\rangle| : \|z\|_x \leq 1\} \quad \text{for all } v \in \mathcal{V}^*. \tag{3.1}$$

For ease of presentation, we tacitly assume in what follows that $\|z\|_x \geq \mu\|z\|$ for some $\mu > 0$ and all $x \in \mathcal{X}$, $z \in \mathcal{Z}$. This can always be achieved by taking $\|\cdot\|_x \leftarrow \|\cdot\|_x + \mu\|\cdot\|$ so there is no loss of generality. Note in particular that this implies that $\|v\|_{x,*} \leq (1/\mu)\|v\|$ for all $x \in \mathcal{X}$ and all $v \in \mathcal{Z}^*$.

For intuition, we present some key examples below:

Example 3.1 (Euclidean geometry)**.** Let $\mathcal{X} = \mathbb{R}^d$ so $\mathcal{Z} = \mathbb{R}^d$. The *Euclidean norm* on $\mathcal{X}$ is given by the standard expression $\|z\|_2^2 = \sum_{j=1}^d z_j^2$, and the associated dual norm is the same.

Example 3.2 (Shahshahani p-norm)**.** Let $\mathcal{X} = \mathbb{R}^d_{++}$ so, again, $\mathcal{Z} = \mathbb{R}^d$. The *Shahshahani p-norm* on $\mathcal{X}$ is defined for all $p > 1$ as

$$\|z\|_x = \left(|z_1|^p/x_1 + \cdots + |z_d|^p/x_d\right)^{1/p} \quad \text{for all } x \in \mathcal{X},\, z \in \mathcal{Z}. \tag{3.2}$$

By a straightforward application of Hölder's inequality, the corresponding dual norm is given by

$$\|v\|_{x,*} = \left(x_1^{q-1}|v_1|^q + \cdots + x_d^{q-1}|v_d|^q\right)^{1/q} \tag{3.3}$$

with the usual convention $p^{-1} + q^{-1} = 1$. In particular, for $p \to 1^+$, we get the limiting expression

$$\|v\|_{x,*} = \max\{x_1|v_1|, \dots, x_d|v_d|\}. \tag{3.4}$$

This metric plays a major role in, among others, game theory, optimal transport, machine learning, information theory, and many other fields – see e.g., [1, 2, 21, 29, 32, 45, 46] and references therein.

Local Bregman functions and the associated divergence. The notion of a dual local norm presented above will be our principal measure of distance in $\mathcal{V}^*$. To proceed, we will also need to adapt the notion of a *Bregman* (or *distance-generating*) *function* on $\mathcal{X}$:

Definition 2. Let $\|\cdot\|_x$ be a local norm on $\mathcal{X}$. We say that $h\colon \mathcal{V} \to \mathbb{R}$ is a *Bregman function* on $\mathcal{X}$ if:

1. h is proper, l.s.c., convex, and $\operatorname{cl}(\operatorname{dom} h) = \operatorname{cl}(\mathcal{X})$.
2. The subdifferential of h admits a *continuous selection*, i.e., a continuous function ∇h such that $\nabla h(x) \in \partial h(x)$ for all $x \in \mathcal{X}^\circ \equiv \operatorname{dom} \partial h$.
3. h is strongly convex relative to the underlying local norm, i.e.,
$$h(p) \geq h(x) + \langle \nabla h(x), p - x\rangle + \tfrac{1}{2}K\|p - x\|_x^2 \tag{3.5}$$
for some $K > 0$ and all $p \in \mathcal{X}$, $x \in \mathcal{X}^\circ$.

The *Bregman divergence* induced by h is then defined for all $p \in \mathcal{X}$, $x \in \mathcal{X}^\circ$ as

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x\rangle. \tag{3.6}$$

As an immediate consequence of the above, we have:

Lemma 2. *A Bregman function h is K-strongly convex relative to $\|\cdot\|_x$ if and only if*

$$D(p, x) \geq \tfrac{1}{2}K\|p - x\|_x^2 \quad \textit{for all } p \in \mathcal{X} \textit{ and all } x \in \mathcal{X}^\circ. \tag{3.7}$$

The main difference between Definition 2 and the standard assumptions in the literature [7, 20, 28, 30, 31, 37–39] is the strong convexity requirement relative to the local norm $\|\cdot\|_x$ (whose choice, in turn, is aimed to capture the singularity landscape of the operator). We illustrate this with two examples below:

[6] By that, we have in mind the definition of an absolutely homogeneous Finsler metric [5]. Specifically, a local norm is viewed here as continuous nonnegative function $F\colon \mathcal{X} \times \mathcal{V} \to \mathbb{R}_+$ with the following propoerties: for all $x \in \mathcal{X}$ and all $z_1, z_2 \in \mathcal{V}$, we have (*i*) $F(x, z_1 + z_2) \leq F(x, z_1) + F(x, z_2)$; (*ii*) $F(x, \lambda z) = |\lambda| z$; and (*iii*) $F(x, z) > 0$ for all $z \in \mathcal{V} \setminus \{0\}$. The local norm of z at x is then defined as $\|z\|_x = F(x, z)$.

Example 3.3. Suppose that $\mathcal{X} = \mathbb{R}^d$ is endowed with the Euclidean norm as in Example 3.1. Then, setting $h(x) = (1/2)\|x\|_2^2$, we get the standard expression $D(p, x) = (1/2)\|p - x\|_2^2$ for the associated Bregman divergence. Obviously, h is 1-strongly convex relative to $\|\cdot\|_2$.

Example 3.4. Let $\mathcal{X} = [0, 1)^d$ (so $\mathcal{X}$ is neither open nor closed), and consider the local norm $\|z\|_x^2 = \sum_{i=1}^d |z|_i^2/(1 - x_i)^2$ for $x \in \mathcal{X}$, $z \in \mathbb{R}^d$ (cf. Example 3.2 above). If we set

$$h(x) = \textstyle\sum_{i=1}^d 1/(1 - x_i) \tag{3.8}$$

a straightforward calculation gives

$$D(p, x) = \sum_{i=1}^d \frac{(p_i - x_i)^2}{(1 - p_i)(1 - x_i)^2} \geq \sum_{i=1}^d \frac{(p_i - x_i)^2}{(1 - x_i)^2} = \|p - x\|_x^2, \tag{3.9}$$

i.e., h is strongly convex relative to $\|\cdot\|_x$. Importantly, since $\|\cdot\|_x \geq \|\cdot\|_2$, this Bregman function is also strongly convex relative to the standard Euclidean norm. However, even though the Euclidean regularizer of Example 3.3 is strongly convex relative to *any* global norm on $\mathcal{X}$, it cannot be strongly convex relative to the local norm $\|\cdot\|_x$ because of the singularity of the latter when $x_i \to 1^-$.

Bregman continuity. We are now in a position to introduce the notion of *Bregman continuity:*

Definition 3. Let h be a local Bregman function relative to some local norm $\|\cdot\|_x$ on $\mathcal{X}$. We say that the operator $A\colon \mathcal{X} \to \mathcal{V}^*$ is β-*Bregman continuous* if

$$\|A(x') - A(x)\|_{x,*} \leq \beta\sqrt{2D(x, x')} \quad \text{for all } x, x' \in \mathcal{X}. \tag{BC}$$

Of course, in the case of a global norm with Bregman function $h(x) = (1/2)\|x\|^2$ (cf. Example 3.3), we recover the standard Lipschitz continuity condition: $\|A(x') - A(x)\|_* \leq \beta\sqrt{2D(x', x)} = \beta\|x' - x\|$. On the other hand, the example below shows that an operator can be Bregman continuous without being Lipschitz continuous relative to *any* global norm:

Example 3.5. Consider the operator $A(x) = (c_r \cdot (1 - x_r/c_r)^{-1})_{r\in\mathcal{R}}$ defined in Example 2.3. Renormalizing c_r to 1 for clarity and using the Bregman data of Examples 3.2 and 3.4, we get:

$$\|A(x') - A(x)\|_{x,*}^2 = \sum_{i=1}^d \frac{(x_i' - x_i)^2}{(1 - x_i')^2} \leq \sum_{i=1}^d \frac{(x_i' - x_i)^2}{(1 - x_i)(1 - x_i')^2} = D(x, x') \tag{3.10}$$

i.e., A is $(1/\sqrt{2})$-Bregman continuous relative to h. However, given the singularity of $A(x)$ as $x_i \to 1^-$, we see that A cannot be Lipschitz continuous relative to *any* global norm on $\mathcal{X}$.

Importantly, this example suggests the following rule of thumb: if the Jacobian of A exhibits a singularity of the form $\mathcal{O}(\phi(x))$ near the residual set $\mathrm{cl}(\mathcal{X}) \setminus \mathcal{X}$ of $\mathcal{X}$, taking $\|\cdot\|_x = \Theta(\phi(x))$ and $h(x) = \Theta(\phi(x))$ allows A to be Bregman continuous, despite this singularity. This heuristic provides a principled choice of Bregman data under which A satisfies (BC).

4 The mirror-prox algorithm

In this section, we present the main algorithmic method that we will use to solve (VI) under Bregman continuity. Our core assumptions in that regard will be:

Assumption 1. The solution set $\mathcal{X}^* \equiv \mathrm{Sol}(\mathcal{X}, A)$ of (VI) is nonempty.

Assumption 2. A is monotone and β-Bregman continuous.

In addition to the above, we assume that the optimizer gains access to A via an *oracle* which, when called at the t-th stage of a sequence $X_t \in \mathcal{X}$, returns (possibly imperfect) feedback of the form

$$V_t = A(X_t) + U_t, \tag{4.1}$$

where $U_t \in \mathcal{V}^*$ is an additive noise variable. The two cases of interest that we consider here are (*i*) when $U_t = 0$ for all t; and (*ii*) when U_t satisfies the statistical hypotheses:

$$a)\ \textit{Zero-mean:} \qquad \mathbb{E}[U_t \mid \mathcal{F}_t] = 0. \tag{4.2a}$$

$$b)\ \textit{Finite variance:} \qquad \mathbb{E}[\|U_t\|_*^2 \mid \mathcal{F}_t] \leq \sigma^2. \tag{4.2b}$$

with $\mathcal{F}_t$ denoting the history (natural filtration) of X_t. For obvious reasons, we will refer to the first case ($U_t = 0$) as a *perfect oracle*, and to the second one as a *stochastic oracle*.

Following Nemirovski [36] and Juditsky et al. [20], the *mirror-prox* (MP) algorithm can be stated in recursive form as follows:

$$\begin{aligned} X_{t+1/2} &= P_{X_t}(-\gamma_t V_t) \\ X_{t+1} &= P_{X_t}(-\gamma_t V_{t+1/2}) \end{aligned} \tag{MP}$$

where $\gamma_t > 0$ is a variable step-size sequence (discussed in detail below), and the so-called "prox-mapping" $P\colon \mathcal{X}^\circ \times \mathcal{V}^* \to \mathcal{X}$ is defined as

$$P_x(y) = \arg\min_{x' \in \mathcal{X}} \{\langle y, x - x' \rangle + D(x', x)\} \tag{4.3}$$

with $D(\cdot,\cdot)$ denoting the divergence of an underlying Bregman function $h\colon \mathcal{X} \to \mathbb{R}$. For concreteness, we also assume in what follows that (MP) is initialized at the so-called "prox-center" of $\mathcal{X}$, i.e.,

$$X_1 = x_c \equiv \arg\min_{x \in \mathcal{X}} h(\mathcal{X}). \tag{4.4}$$

Remark 1. In general, calculating mirror steps can be computationally expensive – just like Euclidean projections in several cases. In what follows, we tacitly assume that our setting is "prox-friendly" [20, 36, 38] in the sense that the update (4.3) can be computed efficiently. For instance, it is straightforward to check that Example 3.4 is prox-friendly.

Heuristically, the main idea behind (MP) is that, at each $t = 1, 2, \dots$, the oracle is called at the algorithm's base state X_t to generate an intermediate, *leading* state $X_{t+1/2}$; subsequently, the base state is updated with oracle information from the leading state $X_{t+1/2}$ and the process repeats. In this way, (MP) essentially tries to "anticipate" the change of A along a prox-step, and to exploit this "forward" information in order to achieve a faster convergence rate than ordinary forward-backward/gradient descent schemes. For this anticipatory scheme to work, the variation of the operator A must be sufficiently gradual, hence the need for Lipschitz continuity in the classical analysis of the algorithm [20, 36, 38]. If this variation is unbounded (e.g., if A exhibits singularities), this look-ahead mechanism could break down completely and the algorithm might fail to converge altogether. Our first result below is that, despite such singularities, Bregman continuity allows us to recover the optimal convergence rate of (MP):

Theorem 1. *Assume that A satisfies Assumptions 1 and 2, and let Gap_H denote the restricted gap function for the Bregman zone $\mathcal{C}_H = \{x \in \mathcal{X} : D(x, x_c) \leq H\}$. Suppose further that (MP) is run with a K-strongly convex Bregman function and oracle feedback of the form (4.1). Then, for all $H > 0$, the averaged sequence $\bar{X}_T = \sum_{t=1}^T \gamma_t X_{t+1/2} \big/ \sum_{t=1}^T \gamma_t$ enjoys the following gap bounds:*

a) If $\sigma^2 = 0$ and the algorithm's step-size satisfies

$$0 < \gamma_{\min} \equiv \inf_t \gamma_t \leq \sup_t \gamma_t \equiv \gamma_{\max} \leq \sqrt{K}/\beta, \tag{4.5}$$

we have

$$\mathrm{Gap}_H(\bar{X}_T) \leq \frac{H}{\gamma_{\min}} \frac{1}{T} \tag{4.6}$$

b) Otherwise, if $\sigma^2 > 0$ and $\gamma_t \leq \sqrt{K/2}/\beta$, we have

$$\mathbb{E}[\mathrm{Gap}_H(\bar{X}_T)] = \mathcal{O}\Big(\frac{H + \sigma^2 \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}\Big) \tag{4.7}$$

In particular, if $\gamma_t \propto 1/\sqrt{T}$, we get $\mathbb{E}[\mathrm{Gap}_H(\bar{X}_t)] = \mathcal{O}(1/\sqrt{T})$.

Remark. We should stress here that the above rates are attained even if $\mathcal{X}$ is unbounded and/or the "Bregman content" $H_{\mathcal{X}} \equiv \sup_x D(x, x_c) = \sup h - \inf h$ of $\mathcal{X}$ is infinite.

As we show in the supplement, the key step in the proof of the deterministic part of Theorem 1 is the following energy inequality for an arbitrary target point $p \in \mathcal{C}_H$:

$$D(p, X_{t+1}) \leq D(p, X_t) - \gamma_t \langle A(X_{t+1/2}), X_{t+1/2} - p \rangle - \Big(1 - \frac{\beta^2 \gamma_t^2}{K}\Big) D(X_{t+1/2}, X_t) \tag{4.8}$$

There are two points where the Bregman structure of the algorithm can be seen in (4.8): in the energy iterates $D(p, X_t)$, but also in the comparison of the algorithm's base and leading state in the term

Algorithm 1: adaptive mirror-prox (AMP)

Require: local norm $\|\cdot\|_x$, K-strongly convex Bregman function h, shrink ratio $\theta \in (0,1)$

take $X_1 = \arg\min h$, $\gamma_1 > 0$ # initialization
for $t = 1, 2, \dots$ **do**
  get oracle feedback V_t at X_t # base state query
  set $X_{t+1/2} = P_{X_t}(-\gamma_t V_t)$ # leading state update
  get oracle feedback $V_{t+1/2}$ at $X_{t+1/2}$ # leading state query
  set $X_{t+1} = P_{X_t}(-\gamma_t V_{t+1/2})$ # base state update
  set $\beta_t = \dfrac{\|V_{t+1/2} - V_t\|_{X_{t+1/2},*}}{\sqrt{2D(X_{t+1/2}, X_t)}}$ # estimate Bregman constant
  set $\gamma_{t+1} = \min\{\gamma_t, \theta\sqrt{K}/\beta_t\}$ # update step-size
end for

$D(X_{t+1/2}, X_t)$. In the "vanilla" setting, Lipschitz continuity is used to obtain a comparison of these successive states in terms of a global norm difference of the form $\|X_{t+1/2} - X_t\|^2$. However, this step also requires A to vary gradually relative to $\|\cdot\|$, which is of course impossible if A exhibits singularities. The key novelty in our setting is the use of the Bregman divergence as a comparator for the algorithm's *successive* states: it is at this point that the triple interplay between the operator, the local norm and the chosen Bregman function is made manifest, and it is what makes Bregman continuity particularly well-suited for tackling singular problems of this kind. This requires a careful treatment of the various Bregman differences involved, so we defer the details to the supplement.

5 The adaptive mirror-prox algorithm

A crucial assumption underlying the analysis of the previous section is that the optimizer must know in advance – or be otherwise able to estimate – the Bregman constant β. In practice, this can be difficult to achieve, so it is important to be able to run (MP) with an *adaptive* step-size policy. Our starting point is the observation that, with accurate feedback, one can estimate β by setting

$$\beta_t = \frac{\|A(X_{t+1/2}) - A(X_t)\|_{X_{t+1/2},*}}{\sqrt{2D(X_{t+1/2}, X_t)}} \tag{5.1}$$

whenever $X_{t+1/2} \neq X_t$; obviously, if A is β-Bregman continuous, we have $\beta_t \leq \beta$. However, the fact that the Bregman constant is being *under*-estimated means that a step-size policy of the form $\gamma_t \propto \sqrt{K}/\beta_t$ would *over*-estimate the inverse Bregman constant $1/\beta$, so the resulting step-size policy would have no reason to satisfy (4.5). To overcome this obstacle, we introduce the following comparison mechanism: first, at each $t = 1, 2, \dots$, we use the estimation (5.1) to test the step-size $\bar{\gamma}_t = \sqrt{K}/\beta_t$. Then, to avoid the growth phenomenon outlined above, we shrink $\bar{\gamma}_t$ by a constant factor of θ and, to avoid running into vanishing step-size issues, we take the previous step-size employed if the shrunk one would be smaller. Formally, we consider the adaptive step-size policy:

$$\gamma_{t+1} = \begin{cases} \min\{\gamma_t, \theta\sqrt{K}/\beta_t\} & \text{if } X_t \neq X_{t+1/2}, \\ \gamma_t & \text{otherwise,} \end{cases} \tag{5.2}$$

with β_t defined as in (5.1) and $\theta \in (0,1)$ chosen arbitrarily.

For concreteness, we call the resulting algorithm *adaptive mirror-prox* (AMP) and we provide a pseudocode implementation in Algorithm 1 above. In terms of performance, we have:

Theorem 2. *Assume that A satisfies Assumptions 1 and 2, and (MP) is run with perfect oracle feedback and the adaptive step-size policy (5.2). Then, with notation as in Theorem 1, the algorithm's ergodic average $\bar{X}_T = \sum_{t=1}^T \gamma_t X_{t+1/2} \big/ \sum_{t=1}^T \gamma_t$ enjoys the gap bound $\mathrm{Gap}_H(\bar{X}_T) = \mathcal{O}(1/T)$.*

We find this result particularly appealing because it yields the optimal $\mathcal{O}(1/T)$ convergence rate of the mirror-prox algorithm, even for possibly singular operators, and even if the operator's Bregman constant is unknown. Its proof relies on using the specific form of the step-size policy (5.2) to control the second term in the energy inequality (4.8); we provide the detailed arguments in the supplement.

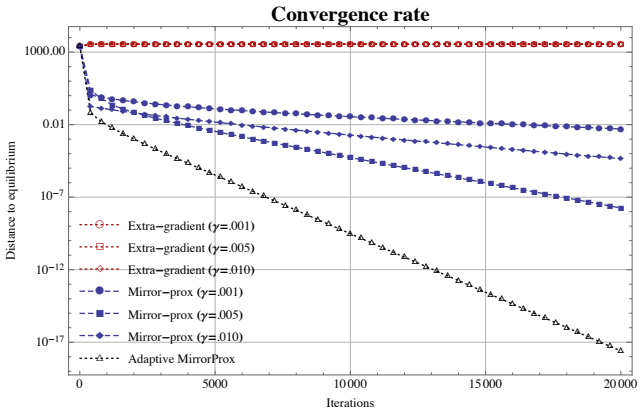

Figure 1: Different variants of the mirror-prox algorithm in the resource sharing problem of Example 2.3. The algorithm labeled "extra-gradient" refers to Euclidean regularization and a constant step size as indicated in the legend; "mirror-prox" was run with the Bregman function of Example 3.4 and step-sizes as in the legend; finally, "adaptive mirror-prox" corresponds to Algorithm 1, i.e., mirror-prox with the adaptive step-size (5.2).

6 Numerical experiments

We performed a series of numerical experiments on the resource sharing problem described in Example 2.3 with a set of $R = 1000$ servers being shared by $N = 100$ commodities, each with a demand drawn uniformly at random from $[0, 1]$; the capacity c_r of each server $r = 1, \dots, R$ was also drawn randomly from $[0, 100]$. Subsequently, we ran two variants of the mirror-prox method: (MP) with Euclidean regularization, and (MP) with the Bregman function defined in Example 3.4. For all methods, we ran a range of different constant step-sizes (we present the most representative values, namely $\gamma = 0.001$, $\gamma = 0.005$, and $\gamma = 0.010$). Subsequently, we also ran Algorithm 1 and we plotted the distance from the solution to the induced variational inequality problem as a function of the number of iterations. The main conclusions that can be drawn are as follows:

1. The Euclidean version of the mirror-prox algorithm (i.e., the extra-gradient algorithm) is unstable and does not converge; this is due to the fact that the gradients received are very large (recall that the problem is *not* Lipschitz continuous), so the algorithm does not exhibit descent or convergence.
2. The MP variant with the non-Euclidean regularizer of Example 3.4 behaves much better and converges (since the VI problem under study is Bregman continuous relative to this Bregman function). However, depending on the method's step-size, the convergence is relatively slow, and there is no easy way to estimate the problem's Bregman constant in order to choose a "good" step-size.
3. By contrast, the adaptive version of the method (Algorithm 1) converges significantly faster than variants with a constant step-size. Importantly, the method's step-size changes significantly from one iteration to the next, at least for the first few thousand iterations, eventually converging to a value close to $\gamma_{\inf} \approx 0.0057$. However, running the method with this constant step-size is significantly slower in the warm-up period of the method, and leads to worse results overall. This is due to the fact that, initially, a greedier step-size is able to take larger steps towards the problem's solution.

7 Concluding remarks

In this work, we introduced a novel regularity condition to account for variational inequalities (both deterministic and stochastic) with possible singularities. This condition, which we call *Bregman continuity*, is tailored to the operator's singularity landscape and, as such, provides the necessary bedrock to achieve optimal convergence rates via a properly chosen version of the mirror-prox algorithm (with or without knowledge of the problem's Bregman constant). This opens up several interesting research directions: First, an appealing extension would be to develop a "model-agnostic" version of the method (which would concurrently provide optimal rates in stochastic and deterministic settings) or to combine it with backtracking / linesearch to accelerate convergence. Finally, it would also be interesting to examine the method's local convergence properties in non-monotone problems (deterministic or stochastic). We relegate these questions to future work.

Acknowledgments

The authors gratefully acknowledge financial support from the French National Research Agency (ANR) under grant ORACLESS (ANR–16–CE33–0004–01) and the COST Action CA16228 "European Network for Game Theory" (GAMENET).

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[2] F. Alvarez, J. Bolte, and O. Brahic. Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, 2004.

[3] F. Bach and K. Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *COLT '19: Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

[4] J.-B. Baillon and G. Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel Journal of Mathematics*, 26:137–150, 1977.

[5] D. D.-W. Bao, S.-S. Chern, and Z. Shen. *An Introduction to Riemann-Finsler Geometry*. Number 200 in Graduate Texts in Mathematics. Springer-Verlag, New York, NY, 2000.

[6] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA, 2 edition, 2017.

[7] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, May 2017.

[8] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3): 2131–2151, 2018.

[9] R. E. Bruck Jr. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1): 159–164, November 1977.

[10] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011.

[11] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, August 1993.

[12] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.

[13] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[14] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

[15] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[16] G. Gidel, R. A. Hemmat, M. Pezehski, R. L. Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

[18] F. Hanzely, P. Richtarik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. https://arxiv.org/abs/1808.03045, 2018.

[19] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

[20] A. Juditsky, A. S. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[21] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13:1865–1890, 2012.

[22] L. Kleinrock. *Queueing Systems*, volume 1: Theory. John Wiley & Sons, New York, NY, 1975.

[23] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

[24] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[25] H. Lu. "Relative-continuity" for non-Lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. https://arxiv.org/abs/1710.04718, 2017.

[26] H. Lu, R. M. Freund, and Y. Nesterov. Relatively-smooth convex optimization by first-order methods and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[27] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

[28] P. Mertikopoulos and W. H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, November 2016.

[29] P. Mertikopoulos and W. H. Sandholm. Riemannian game dynamics. *Journal of Economic Theory*, 177: 315–364, September 2018.

[30] P. Mertikopoulos and M. Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, January 2018.

[31] P. Mertikopoulos and Z. Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, January 2019.

[32] P. Mertikopoulos, E. V. Belmega, R. Negrel, and L. Sanguinetti. Distributed stochastic optimization via matrix exponential learning. *IEEE Trans. Signal Process.*, 65(9):2277–2290, May 2017.

[33] P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

[34] 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 *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[35] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. https://arxiv.org/abs/1901.08511v2, 2019.

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

[37] A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[38] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[39] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1): 221–259, 2009.

[40] N. Nisan, T. Roughgarden, É. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[41] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. https://arxiv.org/abs/1610.09585, October 2016.

[42] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, December 1979.

[43] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

[44] A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013.

[45] S. M. Shahshahani. *A New Mathematical Framework for the Study of Linkage and Selection*. Number 211 in Memoirs of the American Mathematical Society. American Mathematical Society, Providence, RI, 1979.

[46] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. MIT Press, Cambridge, MA, USA, 2012.

[47] G. Stampacchia. Formes bilineaires coercitives sur les ensembles convexes. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 1964.

[48] J. G. Wardrop. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers, Part II*, volume 1, pages 325–378, 1952.

[49] A. Yadav, S. Shah, Z. Xu, D. Jacobs, and T. Goldstein. Stabilizing adversarial nets with prediction methods. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

A Properties of the restricted gap function

In this appendix, we discuss the basic properites of the restricted merit function $\mathrm{Gap}_{\mathcal{C}}$ introduced in the main text. For completeness, we begin with the proof of Lemma 1, itself an extension of a similar result by Nesterov (2007):

Proof of Lemma 1. Let $x^* \in \mathcal{X}$ be a solution of (VI) so $\langle A(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$. Then, by monotonicity, we get:

$$\begin{aligned} \langle A(x), x^* - x \rangle &\leq \langle A(x) - A(x^*), x^* - x \rangle + \langle A(x^*), x^* - x \rangle \\ &= -\langle A(x^*) - A(x), x^* - x \rangle - \langle A(x^*), x - x^* \rangle \leq 0, \end{aligned} \tag{A.1}$$

so $\mathrm{Gap}_{\mathcal{C}}(x^*) \leq 0$. On the other hand, if $x^* \in \mathcal{C}$, we also get $\mathrm{Gap}(x^*) \geq \langle A(x^*), x^* - x^* \rangle = 0$, so we conclude that $\mathrm{Gap}_{\mathcal{C}}(x^*) = 0$.

For the converse statement, assume that $\mathrm{Gap}_{\mathcal{C}}(\hat{x}) = 0$ for some $\hat{x} \in \mathcal{C}$ and suppose that $\mathcal{C}$ contains a neighborhood of $\hat{x}$ in $\mathcal{X}$. First, we claim that the following inequality holds:

$$\langle A(x), x - \hat{x} \rangle \geq 0 \quad \text{for all } x \in \mathcal{C}. \tag{A.2}$$

Indeed, assume to the contrary that there exists some $x_1 \in \mathcal{C}$ such that

$$\langle A(x_1), x_1 - \hat{x} \rangle < 0. \tag{A.3}$$

This would then give

$$0 = \mathrm{Gap}_{\mathcal{C}}(\hat{x}) \geq \langle A(x_1), \hat{x} - x_1 \rangle > 0, \tag{A.4}$$

which is a contradiction. Now, we further claim that $\hat{x}$ is a solution of (VI),i.e.,:

$$\langle A(\hat{x}), x - \hat{x} \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{A.5}$$

Indeed, if we suppose that there exists some $z_1 \in \mathcal{X}$ such that $\langle A(\hat{x}), z_1 - \hat{x} \rangle < 0$, then, due to the continuity of A there exists a neighborhood U' of $\hat{x}$ in $\mathcal{X}$ such that

$$\langle A(x), z_1 - x \rangle < 0 \quad \text{for all } x \in U'. \tag{A.6}$$

Hence, assuming without loss of generality that $U' \subset U \subset \mathcal{C}$ (the latter assumption due to the assumption that $\mathcal{C}$ contains a neighborhood of $\hat{x}$), and taking $\lambda > 0$ sufficiently small so that $x = \hat{x} + \lambda(z_1 - \hat{x}) \in U'$, we get that $\langle A(x), x - \hat{x} \rangle = \lambda \langle A(x), z_1 - \hat{x} \rangle < 0$, in contradiction to (A.2). We conclude that $\hat{x}$ is a solution of (VI), as claimed. □

For intuition, we discuss below the relation of the error function $\mathrm{Gap} \equiv \mathrm{Gap}_{\mathcal{X}}$ to other performance measures that arise in practice (similar considerations also apply to $\mathrm{Gap}_{\mathcal{C}}$ as well):

(1) If $A = \nabla f$ for a convex differentiable function f with $\inf_{x \in \mathcal{X}} f(x) > -\infty$, we have, for all $x \in \mathcal{X}$, $\langle \nabla f(x), \hat{x} - x \rangle \leq f(\hat{x}) - f(x)$, so $\mathrm{Gap}(\hat{x}) \leq \sup_x [f(\hat{x}) - f(x)] \leq f(\hat{x}) - \inf f$.

(2) In saddle-point problems, the quality of a candidate solution $\hat{x} = (\hat{x}_1, \hat{x}_2)$ is often assessed via the *Nikaido–Isoda* (NI) function

$$\mathrm{NI}(\hat{x}) = \sup_{x_2 \in \mathcal{X}_2} f(\hat{x}_1, x_2) - \inf_{x_1 \in \mathcal{X}_1} f(x_1, \hat{x}_2) \tag{NI}$$

provided of course that the right-hand side is well-posed. Similarly to the minimization framework above, if f is convex-concave, we have $\mathrm{Gap}(\hat{x}) \leq \mathrm{NI}(\hat{x})$.

In view of all this, a bound on $\mathrm{Gap}(\hat{x})$ does not immediately translate to a bound on the value of the loss or the Nikaido–Isoda function (for minimization or min-max problems respectively). However, the same arguments used to obtain the convergence rate of an algorithm relative to Gap can be usually adapted to these measures with minimal effort.

B Bregman functions and Bregman continuity

B.1 Properties of Bregman functions

In this appendix, we present some basic facts about Bregman functions and proximal mappings. Similar results exist in the literature in different contexts (see e.g., [20, 38, 39] and references therein), but given that many of our results rely on the use of *local* – as opposed to *global* – norms, we provide here complete statements and proofs.

To begin, we introduce two notions that will be particularly useful in the sequel. The first is the convex conjugate of a Bregman function h, i.e.,

$$h^*(y) = \max_{x \in \mathcal{V}} \{\langle y, x\rangle - h(x)\} \tag{B.1}$$

and the associated primal-dual "mirror map" $Q\colon \mathcal{V}^* \to \mathcal{X}$:

$$Q(y) = \arg\max_{x \in \mathcal{V}} \{\langle y, x\rangle - h(x)\} \tag{B.2}$$

We then have the following basic lemma connecting the above notions:

Lemma B.1. *Let h be a K-strongly convex Bregman function as above. Then, for all $x \in \mathcal{X}^\circ \equiv \operatorname{dom} \partial h$, $y \in \mathcal{V}^*$ we have:*

1. $x = Q(y) \iff y \in \partial h(x)$
2. $x^+ = P_x(y) \iff \nabla h(x) + y \in \partial h(x^+) \iff x^+ = Q(\nabla(x) + y)$
3. *Finally, if $x = Q(y)$ and $p \in \mathcal{X}$, we get:*

$$\langle \nabla h(x), x - p\rangle \leq \langle y, x - p\rangle \tag{B.3}$$

Proof. For the first equivalence, note that x solves (B.1) if and only if $0 \in y - \partial h(x)$ and hence if and only if $y \in \partial h(x)$. Working in the same spirit for the second equivalence, we have that x^+ solves (4.3) if and only if $\nabla h(x) + y \in \partial h(x^+)$ and therefore if and only if $x^+ = Q(\nabla h(x) + y)$. For our last claim, by a simple continuity argument, it is sufficient to show that the inequality holds for the relative interior $\operatorname{ri} \mathcal{X}$ of $\mathcal{X}$. In order to show this, pick a base point $p \in \operatorname{ri} \mathcal{X}$, and let

$$\phi(t) = h(x + t(p - x)) - [h(x) + \langle y, x + t(p - x)\rangle] \quad \text{for all } t \in [0, 1]. \tag{B.4}$$

Since, h is strongly convex and $y \in \partial h(x)$ due to the first equivalence, it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Since, $\psi(t) = \langle \nabla h(x + t(p - x)) - y, p - x\rangle$ is a continuous selection of subgradients of ϕ and both ϕ and ψ are continuous over $[0, 1]$, it follows that ϕ is continuously differentiable with $\phi' = \psi$ on $[0, 1]$. Hence, with ϕ convex and $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0, 1]$, we conclude that $\phi'(0) = \langle \nabla h(x) - y, p - x\rangle \geq 0$ and thus we obtain the result. □

The basic ingredient for establishing connections in the Bregman framework is a generalization of the rule of cosines which is known in the literature as the "three-point identity" [11] and will be the main tool for deriving the main estimations for our analysis. Being more precise, we have the following lemma:

Lemma B.2. *Let h be a Bregman function on $\mathcal{X}$. Then, for all $p \in \mathcal{X}$ and all $x, x' \in \mathcal{X}^\circ$, we have:*

$$D(p, x') = D(p, x) + D(x, x') + \langle \nabla h(x') - \nabla h(x), x - p\rangle \tag{B.5}$$

The proof of this lemma follows as in the classic Bregman case [11] so we omit it and proceed to derive some key bounds for the Bregman divergence before and after a mirror step:

Proposition B.1. *Let h be a local Bregman function with strong convexity modulus $K > 0$. Fix some $p \in \mathcal{X}$ and let $x^+ = P_x(y)$ for some $x \in \mathcal{X}^\circ$ and $y \in \mathcal{V}^*$. We then have:*

$$D(p, x^+) \leq D(p, x) - D(x^+, x) + \langle y, x^+ - p\rangle \tag{B.6}$$

Proof. By the three-point identity established in Lemma B.2, we get:

$$D(p, x) = D(p, x^+) + D(x^+, x) + \langle \nabla h(x) - \nabla h(x^+), x^+ - p\rangle \tag{B.7}$$

By rearranging the terms we get:

$$D(p, x^+) = D(p, x) - D(x^+, x) + \langle \nabla h(x^+) - \nabla h(x), x^+ - p\rangle \tag{B.8}$$

Due to (B.3) and the fact that $x^+ = P_x(y)$ so $\nabla h(x) + y \in \partial h(x^+)$, we get the result. □

Now, using the above estimations, we show below the main inequalities which relate the Bregman divergence between *two* prox-steps:

Proposition B.2. *Let h be a Bregman function on $\mathcal{X}$ and fix some $p \in \mathcal{X}$, $x \in \mathcal{X}^\circ$. Letting $x_1^+ = P_x(y_1)$ and $x_2^+ = P_x(y_2)$, we have:*

$$D(p, x_2^+) \leq D(p, x) + \langle y_2, x_1^+ - p\rangle + [\langle y_2, x_2^+ - x_1^+\rangle - D(x_2^+, x)] \tag{B.9}$$

In particular, we have:

$$D(p, x_2^+) \leq D(p, x) + \langle y_2, x_1^+ - p\rangle + \langle y_2 - y_1, x_2^+ - x_1^+\rangle - D(x_2^+, x_1^+) - D(x_1^+, x). \tag{B.10}$$

Proof. For the first inequality, by applying (B.1), (B.6) for $x_2^+ = P_x(y_2)$, we get:

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) - D(x_2^+, x) + \langle y_2, x_2^+ - p\rangle \\ &= D(p, x) + \langle y_2, x_1^+ - p\rangle + [\langle y_2, x_2^+ - x_1^+\rangle - D(x_2^+, x)] \end{aligned} \tag{B.11}$$

For the second inequality, we need to bound $\langle y_2, x_2^+ - x_1^+\rangle - D_h(x_2^+, x)$. In particular, applying again (B.1), (B.6) for $p = x_2^+$, we get:

$$D(x_2^+, x_1^+) \leq D(x_2^+, x) + \langle y_1, x_1^+ - x_2^+\rangle - D(x_1^+, x) \tag{B.12}$$

and hence:

$$D(x_2^+, x) \geq D(x_2^+, x_1^+) + D(x_1^+, x) + \langle y_1, x_1^+ - x_2^+\rangle. \tag{B.13}$$

So, combining the above inequalities we get:

$$\langle y_2, x_2^+ - x_1^+\rangle - D(x_2^+, x) \leq \langle y_2, x_2^+ - x_1^+\rangle - D(x_2^+, x_1^+) - D(x_1^+, x) - \langle y_1, x_1^+ - x_2^+\rangle \tag{B.14}$$

and thus we get the second inequality as well. □

B.2 Bregman cocoercivity

For comparison to the operator theory literature, we end this section by introducing a notion close to Bregman continuity, that of *Bregman cocoercivity*. In particular, we have the following definition:

Definition 4. Let $A\colon \mathcal{X} \to \mathcal{V}^*$ be an operator, let $\|\cdot\|_x$ be a local norm on $\mathcal{X}$ and let h be a local Bregman function on $\mathcal{X}$ with strong convexity modulus K. We say that A is *δ-Bregman cocoercive* if

$$\langle A(x) - A(x'), x - x'\rangle \geq \delta \, \|A(x) - A(x')\|_{*,x} \|A(x) - A(x')\|_{*,x'} \quad \text{for all } x, x' \in \mathcal{X}. \tag{B.15}$$

This notion is a straightforward extension of ordinary operator cocoercivity, e.g., as defined in [6]. As in the base case, we have:

Lemma B.3. *Let h be a Bregman function with local strong convexity modulus $K > 0$ and let $A\colon \mathcal{X} \to \mathcal{V}^*$ be a δ-Bregman cocoercive operator. Then, A is $1/(\delta\sqrt{K})$-Bregman continuous.*

Proof. By applying the Cauchy-Shwartz inequality to the definition (B.15) of Bregman cocoercivity, we get:

$$\begin{aligned} \delta \|A(x) - A(x')\|_{*,x} \|A(x) - A(x')\|_{*,x'} &\leq \langle A(x') - A(x), x' - x\rangle \\ &\leq \|A(x') - A(x)\|_{*,x'} \|x' - x\|_{x'} \end{aligned} \tag{B.16}$$

Thus, by simplifying and recalling Lemma 2 in the main paper, we get:

$$\|A(x) - A(x')\|_{*,x} \leq \frac{1}{\delta} \|x - x'\|_{x'} \leq \frac{1}{\delta\sqrt{K}} \sqrt{2D(x, x')} \tag{B.17}$$

i.e., A is Bregman continuous with modulus of continuity $1/(\delta\sqrt{K})$, as claimed. □

C Convergence analysis of the mirror-prox algorithm

In this section, we provide the proofs of our main convergence results for the mirror-prox algorithm as presented in Sections 4 and 5 of the main part of the paper.

C.1 Deterministic analysis

The main ingredient of our proof for the deterministic case is the following energy inequality:

Proposition C.1. *Assume that A satisfies Assumption 2 and (MP) is run with perfect oracle feedback. Then, for all $p \in \mathcal{X}$, we have:*

$$D(p, X_{t+1}) \leq D(p, X_t) - \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle - \left(1 - \frac{\beta^2 \gamma_t^2}{K}\right) D(X_{t+\frac{1}{2}}, X_t).$$

Proof. By setting $x = X_t$, $y_1 = -\gamma_t A(X_t)$, $x_1^+ = X_{t+\frac{1}{2}}$, $y_2 = -\gamma_t A(X_{t+\frac{1}{2}})$ and $x_2^+ = X_{t+1}$ in Proposition B.2, we readily obtain:

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &- \gamma_t \langle A(X_{t+\frac{1}{2}}) - A(X_t), X_{t+1} - X_{t+\frac{1}{2}} \rangle \\ &- D(X_{t+1}, X_{t+\frac{1}{2}}) - D(X_{t+\frac{1}{2}}, X_t). \end{aligned} \tag{C.1}$$

Proceeding line-by-line, the Fenchel-Young inequality applied to the function $\phi(x) = \|x\|^2_{X_{t+\frac{1}{2}}}$ further gives

$$\begin{aligned} \langle A(X_{t+\frac{1}{2}}) - A(X_t), X_{t+1} - X_{t+\frac{1}{2}} \rangle \leq &\frac{K}{2\gamma_t} \|X_{t+1} - X_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}}} \\ &+ \frac{\gamma_t}{2K} \|A(X_{t+\frac{1}{2}}) - A(X_t)\|_{X_{t+\frac{1}{2}},*}. \end{aligned} \tag{C.2}$$

Thus, by substituting in (C.1), we get

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &+ \frac{K}{2} \|X_{t+1} - X_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}}} + \frac{\gamma_t^2}{2K} \|A(X_{t+\frac{1}{2}}) - A(X_t)\|^2_{X_{t+\frac{1}{2}},*} \\ &- D(X_{t+1}, X_{t+\frac{1}{2}}) - D(X_{t+\frac{1}{2}}, X_t). \end{aligned} \tag{C.3}$$

and hence, by Lemma 2, we obtain:

$$\begin{aligned} D(p, X_{t+1}) \leq D(p, X_t) &- \gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &+ \frac{\gamma_t^2}{2K} \|A(X_{t+\frac{1}{2}}) - A(X_t)\|^2_{X_{t+\frac{1}{2}},*} - D(X_{t+\frac{1}{2}}, X_t). \end{aligned} \tag{C.4}$$

However, the Bregman continuity of A also yields

$$\|A(X_{t+\frac{1}{2}}) - A(X_t)\|^2_{X_{t+\frac{1}{2}},*} \leq 2\beta^2 D(X_{t+\frac{1}{2}}, X_t) \tag{C.5}$$

so our claim follows by combining Eqs. (C.4) and (C.5). □

We are now in a position to establish the $\mathcal{O}(1/T)$ convergence rate of (MP) for deterministic problems:

Proof of Theorem 1 - deterministic case. Fix some $p \in \mathcal{C}_H$. Since $\gamma_t \leq 1/\beta$ by assumption, a slight rearrangement of Proposition C.1 readily yields:

$$\gamma_t \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_t) - D(p, X_{t+1}) \tag{C.6}$$

Moreover, by the monotonicity of A, we also have:

$$\langle A(p), X_{t+\frac{1}{2}} - p \rangle \leq \langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle. \tag{C.7}$$

Thus, combining the two inequalities above, we get

$$\gamma_t \langle A(p), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_t) - D(p, X_{t+1}) \tag{C.8}$$

and, proceeding to telescope from $t = 1$ to T, we obtain:

$$\sum_{t=1}^{T} \gamma_t \langle A(p), X_{t+\frac{1}{2}} - p \rangle \leq D(p, X_1) - D(p, X_{t+1}) \leq D(p, x_c) \tag{C.9}$$

Then, dividing by $\sum_{t=1}^{T}\gamma_t$ finally yields

$$\langle A(p), \bar{X}_T - p\rangle \leq \frac{D(p,x_c)}{\sum_{t=1}^{T}\gamma_t} \leq \frac{D(p,x_c)}{\gamma_{\min}T}, \tag{C.10}$$

so our result follows by taking the supremum over all $p \in \mathcal{X}$ such that $D(p,x_c) \leq H$ (i.e., over all $p \in \mathcal{C}_H$). □

C.2 Stochastic analysis

We now turn to our results for the convergence of the mirror-prox algorithm in the stochastic case:

Proof of Theorem 1 - stochastic case. Working in the same spirit as for the deterministic case, let $x = X_t$, $y_1 = -\gamma_t V_t$, $x_1^+ = X_{t+\frac{1}{2}}$, $y_2 = -\gamma_t V_{t+\frac{1}{2}}$ and $x_2^+ = X_{t+1}$ in the first part of Proposition B.2. We then get:

$$\begin{aligned}
D(p,X_{t+1}) &\leq D(p,X_t) - \gamma_t\langle V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - p\rangle \\
&\quad + \left[\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t)\right] \\
&\leq D(p,X_t) - \gamma_t\langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle \\
&\quad - \gamma_t\xi_{t+\frac{1}{2}} + \left[\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t)\right]
\end{aligned} \tag{C.11}$$

where we used the feedback decomposition $V_{t+\frac{1}{2}} = A(X_{t+\frac{1}{2}}) + U_{t+\frac{1}{2}}$ for $V_{t+\frac{1}{2}}$ and we set $\xi_{t+\frac{1}{2}} = \langle U_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - p\rangle$ in the last line. By the second part of Proposition B.2, we also have

$$\begin{aligned}
\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t) &\leq \gamma_t\langle V_t - V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle \\
&\quad - D(X_{t+1},X_{t+\frac{1}{2}}) - D(X_{t+\frac{1}{2}},X_t)
\end{aligned} \tag{C.12}$$

Now, by applying the Fenchel-Young inequality to the duality pairing in the above inequality, we get

$$\gamma_t\langle V_t - V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle \leq \frac{\gamma_t^2}{2K}\|V_t - V_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}},*} + \frac{K}{2}\|X_{t+1} - X_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}}}. \tag{C.13}$$

On the other hand, by the stochastic oracle assumption (4.1), we have:

$$\begin{aligned}
\frac{\gamma_t^2}{2K}\|V_t - V_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}},*} &\leq \frac{\gamma_t^2}{K}\|A(X_t) - A(X_{t+\frac{1}{2}})\|^2_{X_{t+\frac{1}{2}},*} + \frac{\gamma_t^2}{K}\|U_t - U_{t+\frac{1}{2}}\|^2_{X_{t+\frac{1}{2}},*} \\
&\leq \frac{2\beta^2\gamma_t^2}{K}D(X_{t+\frac{1}{2}},X_t) + \frac{\gamma_t^2}{\mu K}\|U_t - U_{t+\frac{1}{2}}\|_*^2.
\end{aligned} \tag{C.14}$$

where the last line follows from the Bregman continuity of A (Assumption 2) and the fact that $\|\cdot\|_x \geq \mu\|\cdot\|$ for some $\mu > 0$ and all $x \in \mathcal{X}$ (implying in turn that $\|\cdot\|_{x,*} \leq \mu^{-1}\|\cdot\|_*$ for all $x \in \mathcal{X}$). We thus get

$$\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+1} - X_{t+\frac{1}{2}}\rangle - D(X_{t+1},X_t) \leq \left(\frac{2\beta^2\gamma_t^2}{K} - 1\right)D(X_{t+\frac{1}{2}},X_t) + \frac{\gamma_t^2}{\mu K}\|U_t - U_{t+\frac{1}{2}}\|_*^2 \tag{C.15}$$

Since $\gamma_t^2 \leq K/(2\beta^2)$ by assumption, substituting (C.15) in (C.11) and rearranging yields

$$\begin{aligned}
\gamma_t\langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle &\leq D(p,X_t) - D(p,X_{t+1}) - \gamma_t\xi_{t+\frac{1}{2}} + \frac{\gamma_t^2}{\mu K}\|U_t - U_{t+\frac{1}{2}}\|_*^2 \\
&\leq D(p,X_t) - D(p,X_{t+1})) - \gamma_t\xi_{t+\frac{1}{2}} + \frac{2\gamma_t^2}{\mu K}\left[\|U_t\|_*^2 + \|U_{t+\frac{1}{2}}\|_*^2\right].
\end{aligned} \tag{C.16}$$

In order to bound $\xi_{t+\frac{1}{2}}$, we will need to introduce the auxilliary process

$$Z_{t+1} = \arg\min_{x\in\mathcal{X}}\{\langle U_{t+\frac{1}{2}}, Z_t - x\rangle + \frac{\mu}{\gamma_t}D(x,Z_t)\} \tag{C.17}$$

with $Z_1 = x_c$. We then have

$$-\gamma_t \xi_{t+1} = \gamma_t \langle U_{t+\frac{1}{2}}, p - X_{t+\frac{1}{2}} \rangle = \gamma_t \langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}} \rangle + \gamma_t \langle U_{t+\frac{1}{2}}, p - Z_t \rangle \tag{C.18}$$

In order to bound the term which depends on p, we have the following:

$$\begin{aligned}
\gamma_t \langle U_{t+\frac{1}{2}}, p - Z_t \rangle &= \gamma_t \langle U_{t+\frac{1}{2}}, p - Z_{t+1} \rangle + \gamma_t \langle U_{t+\frac{1}{2}}, Z_{t+1} - Z_t \rangle \\
&\leq \mu \langle \nabla h(Z_{t+1}) - \nabla h(Z_t), p - Z_{t+1} \rangle + \frac{\gamma_t^2}{2K} \|U_{t+\frac{1}{2}}\|_{*,Z_t}^2 + \frac{K}{2} \|Z_{t+1} - Z_t\|_{Z_t}^2 \\
&\leq \mu \langle \nabla h(Z_{t+1}) - \nabla h(Z_t), p - Z_{t+1} \rangle + \frac{\gamma_t^2}{2\mu K} \|U_{t+\frac{1}{2}}\|_*^2 + \frac{K\mu}{2} \|Z_{t+1} - Z_t\|^2.
\end{aligned} \tag{C.19}$$

Hence, by the three-point identity, we obtain:

$$\begin{aligned}
\gamma_t \langle U_{t+\frac{1}{2}}, p - Z_t \rangle &\leq \mu [D(p, Z_t) - D(p, Z_{t+1})] - \mu D(Z_{t+1}, Z_t) \\
&\quad + \frac{\gamma_t^2}{2\mu K} \|U_{t+\frac{1}{2}}\|_*^2 + \frac{K\mu}{2} \|Z_{t+1} - Z_t\|^2 \\
&\leq \mu [D(p, Z_t) - D(p, Z_{t+1})] + \frac{\gamma_t^2}{2\mu K} \|U_{t+\frac{1}{2}}\|_*^2
\end{aligned} \tag{C.20}$$

where the last inequality is a consequence of the strong convexity of h. Thus, combining all these with the fact that A is monotone, we can telescope and obtain

$$\begin{aligned}
\sum_{t=1}^T \gamma_t \langle A(p), X_{t+\frac{1}{2}} - p \rangle &\leq (1+\mu) D(p, x_c) \\
&\quad + \sum_{t=1}^T \gamma_t \langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}} \rangle \\
&\quad + \frac{1}{\mu K} \sum_{t=1}^T \gamma_t^2 \left[2\|U_t\|_*^2 + \frac{5}{2} \|U_{t+\frac{1}{2}}\|_*^2 \right].
\end{aligned}$$

Hence, after dividing by $\sum_{t=1}^T \gamma_t$ and taking the supremum over $p \in \mathcal{C}_H$, we get:

$$\mathrm{Gap}_H(\bar{X}_T) \leq \frac{(1+\mu)H + \sum_{t=1}^T \gamma_t \langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}} \rangle + \frac{1}{\mu K} \sum_{t=1}^T \gamma_t^2 \left[2\|U_t\|_*^2 + \frac{5}{2} \|U_{t+\frac{1}{2}}\|_*^2 \right]}{\sum_{t=1}^T \gamma_t}. \tag{C.21}$$

Since $\mathbb{E}[\langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}} \rangle] = \mathbb{E}[\mathbb{E}[\langle U_{t+\frac{1}{2}}, Z_t - X_{t+\frac{1}{2}} \rangle] \,|\, \mathcal{F}_{t+\frac{1}{2}}] = 0$, taking expectations yields

$$\mathbb{E}[\mathrm{Gap}_H(\bar{X}_t)] \leq \frac{(1+\mu)D + \frac{9\sigma^2}{2\mu K} \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}, \tag{C.22}$$

which proves our claim. Finally, the RHS of this last inequality is $\tilde{\mathcal{O}}(1/T^{1/2})$ if $\gamma_t \propto 1/\sqrt{t}$, so the $\tilde{\mathcal{O}}(1/\sqrt{T})$ result follows. □

C.3 Adaptive analysis

Our aim here is to prove the $\mathcal{O}(1/T)$ convergence rate of the adaptive mirror-prox algorithm:

Proof of Theorem 2. For simplicity, we will assume in what follows that $X_{t+\frac{1}{2}} \neq X_t$ for all $t \geq 1$. Otherwise, if $X_{t+\frac{1}{2}} = X_t$ for some t, we would also have $\gamma_{t+1} = \gamma_t$ for said value of t by convention; this would not change our arguments below, but it would make them much more cumbersome to write down.

With this caveat in mind, we begin by showing that the adaptive step-size policy $\gamma_{t+1} = \min\{\gamma_t, \theta\sqrt{K}/\beta_t\}$ is bounded from below as

$$\gamma_t \geq \min\{\gamma_1, \theta\sqrt{K}/\beta\}. \tag{C.23}$$

We consider two cases below: First, if $\gamma_1 \leq \theta\sqrt{K}/\beta$, we will also have

$$\frac{\theta\sqrt{K}}{\beta_t} \geq \frac{\theta\sqrt{K}}{\beta} \geq \gamma_1 \tag{C.24}$$

for all $t \geq 1$. We then claim that $\gamma_t = \gamma_1$ for all $t \geq 1$: indeed, assuming inductively that this is the case for some $\geq\geq 1$ (the claim is trivially true for $t = 1$), we readily get

$$\gamma_{t+1} = \min\{\gamma_t, \theta\sqrt{K}/\beta_t\} = \min\{\gamma_1, \theta\sqrt{K}/\beta_t\} = \gamma_1, \tag{C.25}$$

which proves our claim in this case.

Assume now that $\gamma_1 > \theta\sqrt{K}/\beta$, in which case we are left to show that $\gamma_t \geq \theta\sqrt{K}/\beta$ for all $t \geq 1$. Again, the base case $t = 1$ being true trivially, assume inductively that this holds for some $t \geq 1$. Then, one of the following will hold:

1. $\gamma_{t+1} = \gamma_t$ so $\gamma_{t+1} \geq \theta\sqrt{K}/\beta$ by the inductive assumption.
2. $\gamma_{t+1} = \theta\sqrt{K}/\beta_t \geq \theta\sqrt{K}/\beta$ by the fact that β_t is an under-estimate of β.

In both cases we obtain $\gamma_{t+1} \geq \theta\sqrt{K}/\beta$, so the induction is complete.

To proceed, going back to the proof of the basic energy inequality (C.1), the intermediate step (C.4) can be rewritten as

$$\begin{aligned} D(p, X_{t+1}) &\leq D(p, X_t) - \gamma_t\langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle - \left(1 - \frac{\beta_t^2\gamma_t^2}{K}\right) D(X_{t+\frac{1}{2}}, X_t) \\ &\leq D(p, X_t) - \gamma_t\langle A(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle - \left(1 - \theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right) D(X_{t+\frac{1}{2}}, X_t), \end{aligned} \tag{C.26}$$

where we used the fact that $\gamma_{t+1} \leq \theta\sqrt{K}/\beta_t$ by construction. However, since γ_{t+1} is weakly decreasing and bounded from below, it follows that γ_t converges to some limit value γ_∞ as $t \to \infty$. In turn, this implies that

$$\lim_{t\to\infty}\left(1 - \theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right) = 1 - \theta^2 > 0, \tag{C.27}$$

and, hence

$$\left(1 - \theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right) D(X_{t+\frac{1}{2}}, X_t) > 0 \tag{C.28}$$

for all t greater than some (finite) t_0. Accordingly, rearranging (C.26) and subsequently telescoping as in (C.9), we finally obtain

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t\langle A(p), X_{t+\frac{1}{2}} - p\rangle &\leq D(p, X_1) - D(p, X_{t+1}) - \sum_{t=1}^{T}\left(1 - \theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right) D(X_{t+\frac{1}{2}}, X_t) \\ &\leq D(p, x_c) - \sum_{t=1}^{t_0}\left(1 - \theta^2\frac{\gamma_t^2}{\gamma_{t+1}^2}\right) D(X_{t+\frac{1}{2}}, X_t) < \infty \end{aligned} \tag{C.29}$$

whenever $T > t_0$. Our result then follows by dividing both sides of this last inequality by $\sum_{t=1}^{T}\gamma_t$ and recalling the fact that $\gamma_t \geq \min\{\gamma_1, \theta\sqrt{K}/\beta\} > 0$ for all t. □