ON THE RATE OF CONVERGENCE OF BREGMAN PROXIMAL METHODS IN CONSTRAINED VARIATIONAL INEQUALITIES

WAÏSS AZIZIAN[*,♯], FRANCK IUTZELER[*],
JÉRÔME MALICK[◇], AND PANAYOTIS MERTIKOPOULOS[§]

Abstract. We examine the last-iterate convergence rate of Bregman proximal methods – from mirror descent to mirror-prox – in constrained variational inequalities. Our analysis shows that the convergence speed of a given method depends sharply on the *Legendre exponent* of the underlying Bregman regularizer (Euclidean, entropic, or other), a notion that measures the growth rate of said regularizer near a solution. In particular, we show that boundary solutions exhibit a clear separation of regimes between methods with a zero and non-zero Legendre exponent respectively, with linear convergence for the former versus sublinear for the latter. This dichotomy becomes even more pronounced in linearly constrained problems where, specifically, Euclidean methods converge along sharp directions in a finite number of steps, compared to a linear rate for entropic methods.

1. Introduction

This paper focuses on solving variational inequality (VI) problems of the form

$$\text{Find } x^* \in \mathcal{X} \text{ such that } \langle v(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{X}, \tag{VI}$$

where $\mathcal{X}$ is a closed convex subset of a finite-dimensional normed space $\mathcal{V}$, and $v\colon \mathcal{X} \to \mathcal{V}^*$ is a single-valued operator on $\mathcal{X}$ with values in $\mathcal{V}^*$, the dual of $\mathcal{V}$. The study of such problems dates back to Stampacchia [39] and Minty [24], and it has recently attracted considerable interest in many areas of mathematical programming, game theory and data science as a template for "optimization beyond minimization" – i.e., for problems where finding an optimal state does not necessarily involve minimizing a loss function. In particular, in addition to standard minimization problems – which are recovered when $v = \nabla f$ for some smooth function f – the general formulation (VI) includes saddle-point problems, games, complementarity problems, systems of nonlinear equations, and many other types of equilibrium problems; for a comprehensive introduction to the topic and its applications, see Facchinei & Pang [11] and references therein.

Algorithms for solving (VI) likewise have a rich history in optimization. To provide a short overview, the original proximal point methods [20, 37] were shown to converge when v is monotone; however, these methods involve a backward step on v, so they are difficult to implement in practice. In this regard, forward methods that only require oracle access to v are more practical, but they fail to converge if v is merely monotone. Nonetheless, if

[*] Univ. Grenoble Alpes, LJK, Grenoble 38000, France.
[♯] DI, ENS, Univ. PSL, 75005, Paris, France.
[◇] Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.
[§] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France.
2020 *Mathematics Subject Classification.* Primary 65K15, 90C33; secondary 68Q25, 68Q32.
Key words and phrases. Legendre exponent; optimistic mirror descent; variational inequalities.

coupled with an iterate averaging scheme, forward methods *do* converge, and they achieve an $\mathcal{O}(1/\sqrt{t})$ convergence certificate after t iterations [9, 32].

Bridging the gap between forward and backward methods, the extra-gradient (EG) algorithm [17] provided an extrapolation mechanism that emulated a backward step with two interleaved forward steps, achieving in this way trajectory convergence when v is pseudomonotone and Lipschitz continuous. Subsequently, combining extrapolation with iterate averaging, the mirror-prox (MP) algorithm of Nemirovski [27] – which has the same update structure as [4] – was shown to converge at a rate of $\mathcal{O}(1/t)$ in monotone VIs, a rate which is order-optimal for first-order methods [26, 31].

A salient feature of the mirror-prox algorithm is that it foregoes Euclidean projections in favor of a more sophisticated Bregman proximal step in the spirit of mirror descent (MD) [4, 7, 28]. Owing to this feature, mirror-prox achieves an almost dimension-free convergence speed in problems with a favorable geometry, all the while retaining an order-optimal dependence on t. Because of this "best-of-both-worlds" guarantee, the mirror-prox algorithm and its variants – dualized [30], optimistic [14, 34, 35], stochastic [12, 16], adaptive [2, 3], etc. – have become the method of choice for large-scale variational inequality problems where higher-order methods cannot be efficiently implemented.

Our contributions. In this broad context, our paper seeks to quantify the finer trajectory convergence properties of Bregman proximal methods as a function of the geometry of the problem and the Bregman regularizer underlying the method. For generality, we focus on non-monotone VIs, and we examine the rate of convergence of a wide class of *abstract mirror-prox* (AMP) algorithms to local solutions that satisfy a second-order sufficient condition. Specifically, the AMP template includes as special cases the mirror-prox, mirror descent and optimistic mirror descent algorithms, so it provides a unified view of some of the most widely used Bregman methods in the literature.

Our first finding is that the algorithm's rate of convergence depends sharply on the chosen Bregman regularizer (Euclidean, entropic, or other). We formalize this via the notion of the *Legendre exponent*, which can roughly be described as the logarithmic ratio of the volume of a ball centered at the solution under study to that of a Bregman ball of the same radius.[1] For example, Euclidean methods have a Legendre exponent of $\beta = 0$ and they converge at a linear rate; entropic methods have a Legendre exponent of $\beta = 1/2$ at boundary points, and they converge at a rate of $\mathcal{O}(t^{-1})$; and more generally, as we show in Theorem 1, methods with a Legendre exponent $\beta > 0$ converge at a rate of $\mathcal{O}(t^{1-1/\beta})$. The Euclidean regime ($\beta = 0$) is perfectly aligned with existing results for the geometric last-iterate convergence rate of the EG algorithm and its variants in strongly monotone VIs [12, 14, 19, 25]. By contrast, the Legendre regime ($\beta > 0$) indicates a significant drop in the algorithm's last-iterate convergence speed, even though ergodic convergence results might suggest otherwise.

Subsequently, we take a closer look at the convergence rate of AMP methods as a function of the constraints that are active at a solution x^* of (VI) and the position of $v(x^*)$ relative to said constraints. This analysis reveals that Bregman proximal methods have a particularly fine structure: along *sharp directions* (i.e., constraints along which $v(x^*)$ is strictly inward-pointing), AMP algorithms converge in a finite number of iterations if $\beta = 0$, at a geometric rate if $0 < \beta \leq 1/2$, and at a rate of $\mathcal{O}(1/t^{1/(2\beta-1)})$ if $1/2 < \beta < 1$ (cf. Theorem 2 for a precise statement). Thus, even though the rate estimates of Theorem 1 are in general tight,

[1] This notion was first introduced in the conference paper [5], which can be seen as a precursor of our work. The paper [5] deals with *stochastic* variational inequalities (so there is no overlap with the results presented herein and the derived rates are naturally different), but the notion of the Legendre exponent plays a similar role in both works.

the actual rate of convergence of a Bregman method along different coordinates / constraints could be considerably different – and, in fact, dramatically faster if the solution under study is itself sharp. We find this separation property particularly appealing, as it highlights the interplay between sharp and non-sharp directions: Theorem 1 describes the rate of convergence along the latter, while Theorem 2 estimates the speed along the former.

2. Problem setup and preliminaries

2.1. **Notation.** In the rest of our paper, $\mathcal{V}$ will denote a n-dimensional real space with norm $\|\cdot\|$ and $\mathcal{X}$ will be a closed convex subset thereof. We will also write $\mathcal{Y} \coloneqq \mathcal{V}^*$ for the dual of $\mathcal{V}$, $\langle y, x\rangle$ for the canonical pairing between $y \in \mathcal{Y}$ and $x \in \mathcal{V}$, and $\|y\|_* \coloneqq \max\{\langle y, x\rangle : \|x\| \leq 1\}$ for the induced dual norm on $\mathcal{Y}$. In addition, if $f\colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is an extended-real-valued convex function on $\mathcal{V}$, we will write $\operatorname{dom} f = \{x \in \mathcal{V} : f(x) < \infty\}$ for its effective domain, $\partial f(x) = \{y \in \mathcal{Y} : f(x') \geq f(x) + \langle y, x' - x\rangle \text{ for all } x' \in \mathcal{V}\}$ for the subdifferential of f at $x \in \mathcal{V}$, and $\operatorname{dom} \partial f = \{x \in \mathcal{V} : \partial f(x) \neq \varnothing\}$ for the domain of subdifferentiability of f. Finally, we will make frequent use of Landau's asymptotic notation, writing in particular (*i*) $f(t) = \mathcal{O}(g(t))$ and $g(t) = \Omega(f(t))$ when $\limsup_{t\to\infty} f(t)/g(t) < \infty$; (*ii*) $f(t) = \Theta(g(t))$ when $f(t) = \mathcal{O}(g(t))$ and $f(t) = \Omega(g(t))$; (*iii*) $f(t) = o(g(t))$ when $\limsup_{t\to\infty} f(t)/g(t) = 0$; and (*iv*) $f(t) \sim g(t)$ when $\lim_{t\to\infty} f(t)/g(t) = 1$.

2.2. **Problem statement and basic assumptions.** As we mentioned in the introduction, we will focus throughout on solving variational inequalities of the form:

$$\text{Find } x^* \in \mathcal{X} \text{ such that } \langle v(x^*), x - x^*\rangle \geq 0 \text{ for all } x \in \mathcal{X} \tag{VI}$$

where $v\colon \mathcal{X} \to \mathcal{Y}$ is a single-valued operator, which we call the problem's *defining vector field*, and for which we make the following blanket assumption:

Assumption 1 (Lipschitz continuity)**.** v is *L-Lipschitz continuous*, i.e.,

$$\|v(x') - v(x)\|_* \leq L\|x' - x\| \quad \text{for all } x, x' \in \mathcal{X}. \tag{LC}$$

For concreteness, we provide below two archetypal examples of VI problems of this type:

Example 2.1 (Function minimization)**.** Consider the minimization problem

$$\min_{x\in\mathcal{X}} f(x) \tag{Opt}$$

with $f\colon \mathcal{X} \to \mathbb{R}$ assumed smooth. Then, letting $v(x) = \nabla f(x)$, the solutions of (VI) are precisely the Karush–Kuhn–Tucker (KKT) points of (Opt) [11]. $\lozenge$

Example 2.2 (Saddle-point problems)**.** A *saddle-point* – or *min-max* – problem can be stated in normal form as

$$\min_{x_1\in\mathcal{X}_1} \max_{x_2\in\mathcal{X}_2} \mathcal{L}(x_1, x_2) \tag{SP}$$

where $\mathcal{X}_1 \subseteq \mathbb{R}^{n_1}$ and $\mathcal{X}_2 \subseteq \mathbb{R}^{n_2}$ are convex and closed, and the problem's objective function $\mathcal{L}\colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ is again assumed to be smooth. In the game-theoretic interpretation of the problem, x_1 is controlled by a player seeking to minimize $\mathcal{L}(\cdot, x_2)$, whereas x_2 is controlled by a player seeking to maximize $\mathcal{L}(x_1, \cdot)$. Accordingly, solving (SP) consists of finding a *Nash equilibrium* (NE), i.e., an action profile $(x_1^*, x_2^*) \in \mathcal{X} \coloneqq \mathcal{X}_1 \times \mathcal{X}_2$ such that $\mathcal{L}(x_1^*, x_2) \leq \mathcal{L}(x_1^*, x_2^*) \leq \mathcal{L}(x_1, x_2^*)$ for all $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$. If $\mathcal{L}$ is not convex-concave, Nash equilibria may fail to exist, in which case one typically looks for *first-order stationary* (FOS) points of $\mathcal{L}$, i.e., action profiles $(x_1^*, x_2^*) \in \mathcal{X}_1 \times \mathcal{X}_2$ such that x_1^* is a KKT point of $\mathcal{L}(\cdot, x_2^*)$ and, respectively, x_2^* is a KKT point of $-\mathcal{L}(x_1^*, \cdot)$. In this case, if we set $x = (x_1, x_2)$ and

$v = (\nabla_{x_1}\mathcal{L}, -\nabla_{x_2}\mathcal{L})$, it is straightforward to check that the solutions of (VI) are precisely the first-order stationary points of $\mathcal{L}$. $\lozenge$

The above examples show that not all solutions of (VI) are desirable: for example, in the case of (Opt), such a solution could be a local *maximum* of f. For this reason, we will concentrate on solutions x^* of (VI) that satisfy the following sufficiency condition:

Assumption 2 (Second-order sufficiency)**.** There exists a convex neighborhood $\mathcal{B}$ of x^* in $\mathcal{X}$ and a positive constant $\mu > 0$ such that

$$\langle v(x) - v(x^*), x - x^* \rangle \geq \mu \|x - x^*\|^2 \quad \text{for all } x \in \mathcal{B}. \tag{SOS}$$

In general, Assumption 2 guarantees that x^* is the unique solution of (VI) in $\mathcal{B}$. In particular, in the setting of (Opt), Assumption 2 implies that f grows (at least) quadratically along every ray emanating from x^*, i.e., $f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle + (\mu/2)\|x - x^*\|^2 = \Omega(\|x - x^*\|^2)$ for all $x \in \mathcal{B}$ (though, of course, this does not mean that f is strongly convex in $\mathcal{B}$). Analogously, for (SP), Assumption 2 gives $\mathcal{L}(x_1, x_2^*) = \Omega(\|x_1 - x_1^*\|^2)$ and $\mathcal{L}(x_1^*, x_2) = -\Omega(\|x_2 - x_2^*\|^2)$, so x^* is a local Nash equilibrium of $\mathcal{L}$. In view of the above, we will focus throughout on solutions satisfying (SOS).

2.3. **Bregman proximal methods.** The algorithmic framework that we will consider is a general class of Bregman proximal methods that we collectively refer to as the *abstract mirror-prox* template. The principal ingredient of these methods is that of a *Bregman regularizer* – or *distance-generating function* (DGF) – which we define below as follows:

Definition 1 (Bregman regularizers and related notions)**.** A proper, lower semi-continuous, strictly convex function $h\colon \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ is said to be a *Bregman regularizer* on $\mathcal{X}$ if

(1) h is supported on $\mathcal{X}$, i.e., $\operatorname{dom} h = \mathcal{X}$.
(2) The subdifferential of h admits a *continuous selection*, i.e., there exists a continuous mapping $\nabla h\colon \operatorname{dom} \partial h \to \mathcal{Y}$ such that $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom} \partial h$.
(3) h is 1-strongly convex relative to $\|\cdot\|$, i.e., for all $x \in \operatorname{dom} \partial h$, $x' \in \operatorname{dom} h$, we have

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2. \tag{1}$$

For posterity, the set $\mathcal{X}_h \coloneqq \operatorname{dom} \partial h$ will be referred to as the *prox-domain* of h. We also define the *Bregman divergence* of h as

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle \qquad \text{for all } x \in \mathcal{X}_h,\ p \in \mathcal{X} \tag{2}$$

and the induced *Bregman proximal mapping* as

$$P_x(y) = \operatorname*{arg\,min}_{x' \in \mathcal{X}} \{\langle y, x - x' \rangle + D(x', x)\} \qquad \text{for all } x \in \mathcal{X}_h,\ y \in \mathcal{Y}. \tag{3}$$

Example 2.3. A staple choice for h is the Euclidean regularizer $h(x) = \frac{1}{2}\|x\|_2^2$. This choice gives

$$D(p, x) = \tfrac{1}{2}\|p - x\|^2 \quad \text{and} \quad P_x(y) = \Pi_{\mathcal{X}}(x + y), \tag{4}$$

with $\Pi_{\mathcal{X}}(x) \coloneqq \arg\min_{x' \in \mathcal{X}} \|x' - x\|$ denoting the Euclidean projector on $\mathcal{X}$. $\lozenge$

Further examples of Bregman regularizers are given in Section 3, where we also take an in-depth look into their properties. For now, given a Bregman regularizer on $\mathcal{X}$, we proceed to define the *abstract mirror-prox* (AMP) template via the generic recursion

$$x_{t+1/2} = P_{x_t}(-\gamma_t V_t) \qquad x_{t+1} = P_{x_t}(-\gamma_t V_{t+1/2}) \tag{AMP}$$

where

(1) $t = 1, 2, \dots$ denotes the method's iteration counter.

(2) $\gamma_t > 0$ is a (non-increasing) step-size sequence.

(3) V_t and $V_{t+1/2}$ are sequences of "oracle signals" whose precise definition we discuss below.

In terms of vocabulary (and for reasons that will also become clear in the sequel), the iterates x_t, $t = 1, 2, \dots$, will be referred to as the "*base states*" of the method, while the "half-iterates" $x_{t+1/2}$, $t = 1, 2, \dots$, will be referred to as the method's "*leading states*". Finally, in terms of initialization, we will take for convenience $x_1 = x_{1/2}$.

Now, to link (AMP) to the problem under study, we will assume throughout that the sequence of oracle signals $V_{t+1/2}$ is generated by querying v at $x_{t+1/2}$, i.e.,

$$V_{t+1/2} = v(x_{t+1/2}) \quad \text{for all } t = 1, 2, \dots \tag{5}$$

In words, (AMP) generates a new base state x_{t+1} by taking a Bregman proximal step from x_t with oracle information from the leading state $x_{t+1/2}$. By contrast, $x_{t+1/2}$ can be generated in a number of different ways, depending on the definition of V_t. We present three prototypical examples below:

(1) The *mirror-prox* (MP) update:

$$V_t = v(x_t) \qquad\qquad \text{for all } t = 1, 2, \dots \tag{MP}$$

The motivation behind (MP) is that the algorithm tries to make more informed steps towards a solution of (VI) by "anticipating" the change of v via a second oracle query. In the context of the Euclidean regularizer (4), the recursion (MP) is known as the *extra-gradient* (EG) algorithm, and was originally proposed by Korpelevich [17]; for the bona fide Bregman version of the algorithm (and namesake of the method), see Nemirovski [27] and Juditsky et al. [16].

(2) The *mirror descent* (MD) update is defined as

$$V_t = 0 \qquad\qquad \text{for all } t = 1, 2, \dots \tag{MD}$$

In this case, the method foregoes any look-ahead efforts and proceeds in a series of Bregman proximal steps $x_{t+1} \leftarrow P_{x_t}(-\gamma_t v(x_t))$. This method has a long history dating back to Nemirovski & Yudin [28]; for an appetizer, we refer the reader to [7, 29, 40, 42] and references therein.

(3) The *optimistic mirror descent* (OMD) update:

$$V_t = v(x_{t-1/2}) \qquad \text{for all } t = 1, 2, \dots \tag{OMD}$$

The idea of this update is to lighten the per-iteration complexity of (MP) by making only a *single* query to v: this is done by approximating $v(x_{t+1/2})$ with the previously available oracle signal, i.e., taking $V_{t+1/2} \leftarrow v(x_{t-1/2})$. This "oracle reuse" idea dates back to Popov [34], and it has been subsequently popularized in machine learning and other fields by [10], [35], and many others; for an overview, see [14] and references therein.

The three algorithms described above are the most widely studied Bregman methods in the literature, so we will use them as running examples throughout. More generally, we will make the following assumption for the input signal V_t generated at the method's base state.

Assumption 3. For all $t = 1, 2, \dots$, the oracle signal V_t is of the form:

$$V_t = av(x_t) + bv(x_{t-1/2}) \tag{6}$$

for some $a, b \in [0, 1]$ with $a + b \leq 1$ and $a + b = 1$ if $b > 0$.

In the language of Assumption 3, the three examples of (AMP) above can be recovered as follows:

- For (MP): $a = 1$, $b = 0$.
- For (MD): $a = 0$, $b = 0$.
- For (OMD): $a = 0$, $b = 1$.

More general input sequences can also be considered – e.g., to cover for sequential or "k-to-1" update rules that are sometimes used in min-max problems [15] – but this would complicate the notation and the resulting rates, so Assumption 3 will suffice for our purposes. We also note in passing that the requirement "$a + b = 1$ if $b > 0$" is only introduced to ease notation and does not lead to a loss in generality: if $b > 0$, we can always rescale the method's step-size by $a + b$ so that the condition $a + b = 1$ is satisfied automatically.

For future use, we close this section with some basic properties of the Bregman divergence and the induced proximal mapping:

Lemma 1. *Let h be a Bregman regularizer on $\mathcal{X}$ and let ∇h be a continuous selection of ∂h. Then, for all $x \in \mathcal{X}_h$, $x^+ \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have:*

a) $\partial h(x) = \nabla h(x) + \mathrm{PC}(x)$ (7a)

b) $x^+ = P_x(y) \iff \nabla h(x) + y \in \partial h(x^+) \iff \nabla h(x^+) - \nabla h(x) \in y - \mathrm{PC}(x^+)$ (7b)

where $\mathrm{PC}(x) = \{y \in \mathcal{Y} : \langle y, x' - x\rangle \leq 0$ for all $x' \in \mathcal{X}\}$ denotes the polar cone to $\mathcal{X}$ at x. In particular, (AMP) *is well-posed: $x^+ = P_x(y)$ implies that $x^+ \in \mathcal{X}_h$.*

Lemma 2 (3-point identity)**.** *For all $p \in \mathcal{X}$ and all $x, x^+ \in \mathcal{X}_h$, we have:*

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

Lemma 3 (Non-expansiveness)**.** *For all $x \in \mathcal{X}_h$ and all $y, y^+ \in \mathcal{Y}$ we have:*

$$\|P_x(y^+) - P_x(y)\| \leq \|y^+ - y\|_* \tag{9}$$

These properties are fairly well known, so we omit their proofs; for a detailed treatment, we defer the interested reader to Beck & Teboulle [7], Juditsky et al. [16], and Mertikopoulos & Zhou [22].

3. Motivating examples

In this section, we take a closer look at some commonly used Bregman regularizers and the induced prox-mappings with the goal of determining the rate of convergence of the associated Bregman proximal method. For concreteness, we will focus on one-dimensional problems where $\mathcal{X}$ is a closed interval of $\mathbb{R}$ (possibly infinite), and v is the affine operator

$$v(x) = x - x^*, \quad x \in \mathbb{R}, \tag{10}$$

for different choices of $x^* \in \mathbb{R}$. Moreover, for ease of presentation and notation, we will only examine the mirror descent recursion (MD) with constant step-size schedules of the form $\gamma_t \equiv \gamma$ for some $\gamma > 0$. In this case, we obtain the general recursive scheme

$$x_{t+1} = F(x_t) \quad \text{with} \quad F(x) = P_x(-\gamma v(x)), \tag{11}$$

and we will examine the rate of convergence of x_t to x^* by analyzing the behavior of F near x^*.

Example 3.1 (Euclidean regularization)**.** We begin with the quadratic regularizer $h(x) = x^2/2$ for $x \in \mathcal{X}$. Concretely, taking $\mathcal{X} = [0, \infty)$ and noting that $h'(x) = x$, we have:

$$\begin{aligned}
&a)\ \text{Prox-domain:} && \mathcal{X}_h = \mathcal{X} \\
&b)\ \text{Bregman divergence:} && D(p, x) = (p - x)^2/2 \\
&c)\ \text{Prox-mapping:} && P_x(y) = [x + y]_+
\end{aligned} \tag{12}$$

Consider now the case $x^* = 0$, i.e., $v(x) = x$. Then, for $\gamma \in (0, 1)$, the update (11) becomes

$$F(x) = x - \gamma x = (1 - \gamma)x \tag{13}$$

i.e., F is contracting. We thus conclude that x_t converges to $x^* = 0$ at a geometric rate, viz.

$$D(x^*, x_t) = \tfrac{1}{2}x_t^2 = \Theta\big((1 - \gamma)^{2t}\big) \text{ or, in absolute value, } |x_t - x^*| = \Theta((1 - \gamma)^t). \qquad \Diamond$$

Example 3.2 (Entropic regularization)**.** Another popular choice when $\mathcal{X} = [0, \infty)$ is the entropic regularizer $h(x) = x \log x$ [6, 7, 38]. In this case, we have $h'(x) = 1 + \log x$, which gives the following:

$$\begin{aligned}
&a)\ \text{Prox-domain:} && \mathcal{X}_h = \operatorname{ri} \mathcal{X} = (0, \infty) \\
&b)\ \text{Bregman divergence:} && D(p, x) = p \log(p/x) + x - p \\
&c)\ \text{Prox-mapping:} && P_x(y) = x \exp(y)
\end{aligned} \tag{14}$$

Now, taking $v(x) = x$ as in the previous example, the update rule (11) becomes

$$F(x) = x \exp(-\gamma x) = x(1 - \gamma x + o(x)) = x - \gamma x^2 + o(x^2) \quad \text{as } x \to 0. \tag{15}$$

In contrast to (13), we now have $F(x) \sim x$ instead of $(1-\gamma)x$, so F is no longer a contraction. Instead, the iterates of (15) may be analyzed by means of the following lemma.

Lemma 4. *Suppose that $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ admits the asymptotic expansion*

$$f(x) = x - \lambda x^{1+r} + o(x^{1+r}) \quad \text{as } x \to 0 \tag{16}$$

for positive constants $\lambda, r > 0$. Then, for $u_1 > 0$ small enough, the sequence $u_{t+1} = f(u_t)$, $t = 1, 2, \dots,$ converges to 0 at a rate of $u_t \sim (\lambda r t)^{-1/r}$.

By means of this lemma (which we prove in Appendix A), we conclude that the iterates of (15) converge to 0 at a rate of

$$D(x^*, x_t) = x_t = |x_t - x^*| \sim 1/(\gamma t). \qquad \Diamond$$

Example 3.3 (Fractional power)**.** Take $\mathcal{X} = [0, \infty)$ and $v(x) = x$ as in Examples 3.1 and 3.2 above. Then, for a given $q > 0$, $q \neq 1$, the *fractional power* regularizer – or *Tsallis entropy* – on $\mathcal{X}$ is defined as $h(x) = [q(1 - q)]^{-1}(x - x^q)$ [1, 21, 41]. For this choice of regularizer, we have $h'(x) = (1 - qx^{q-1})/[q(1 - q)]$, and a series of straightforward calculations gives the following:[2]

$$
\begin{aligned}
&a) \text{ Prox-domain:} && \mathcal{X}_h = (0,\infty) \text{ if } q \in (0,1) \text{ and } \mathcal{X}_h = [0,\infty) \text{ if } q > 1 \\
&b) \text{ Bregman divergence:} && D(p,x) = \frac{x^q - p^q}{q(1-q)} - x^{q-1}\frac{x-p}{1-q} \\
&c) \text{ Prox-mapping:} && P_x(y) = \left[x^{q-1} - (1-q)y\right]^{\frac{1}{q-1}} \quad \text{for } q \in (0,1)
\end{aligned}
\tag{17}
$$

Now, when applied to $v(x) = x$, the fractional power variant of (11) for $q \in (0,1)$ gives

$$F(x) = x\,[1 + \gamma(1-q)x^{2-q}]^{1/(q-1)} = x - \gamma x^{3-q} + o(x^{3-q}) \quad \text{as } x \to 0. \tag{18}$$

Hence, by Lemma 4, we conclude that x_t converges to 0 at a rate of

$$D(x^*, x_t) = \Theta\big(t^{-q/(2-q)}\big) \text{ or, in absolute value, } |x_t - x^*| = \Theta\big(t^{-1/(2-q)}\big). \qquad \lozenge$$

Example 3.4 (Hellinger distance)**.** Our last example concerns the Hellinger regularizer $h(x) = -\sqrt{1-x^2}$ on $\mathcal{X} = [-1,1]$. Since $h'(x) = x/\sqrt{1-x^2}$, we readily obtain the following:

$$
\begin{aligned}
&a) \text{ Prox-domain:} && \mathcal{X}_h = \operatorname{ri}\mathcal{X} = (-1,1) \\
&b) \text{ Bregman divergence:} && D(p,x) = \frac{1 - px - \sqrt{(1-p^2)(1-x^2)}}{\sqrt{1-x^2}} \\
&c) \text{ Prox-mapping:} && P_x(y) = \frac{x + y\sqrt{1-x^2}}{\sqrt{1 - x^2 + (x + y\sqrt{1-x^2})^2}}
\end{aligned}
\tag{19}
$$

In this case, taking $v(x) = x$ as per the previous examples, yields

$$F(x) = \frac{x - \gamma x\sqrt{1-x^2}}{\sqrt{1 - x^2 + (x - \gamma x\sqrt{1-x^2})^2}} \sim x - \gamma x \quad \text{as } x \to 0, \tag{20}$$

i.e., x_t converges to $x^* = 0$ at a geometric rate, as in Example 3.1. On the other hand, if we consider the shifted operator $v(x) = x + 1$, a somewhat tedious calculation (which we detail in Appendix A) gives the following Taylor expansion near $x^* = -1$:

$$F(x) = x^* + (x - x^*) - 2\sqrt{2}\gamma(x - x^*)^{5/2} + o\Big((x - x^*)^{5/2}\Big). \tag{21}$$

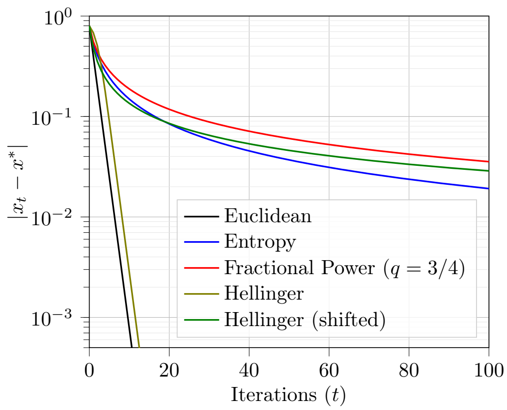
Hence, by Lemma 4, we conclude that x_t converges to $x^* = -1$ at a rate of

$$D(x^*, x_t) = \Theta(t^{-1/3}) \text{ or, in absolute value, } |x_t - x^*| = \Theta(t^{-2/3}). \qquad \lozenge$$

Albeit one-dimensional, the above examples provide a representative view of the geometry of Bregman proximal methods near a solution. Specifically, they show that the Bregman divergence induced by a given regularizer may exhibit a drastically different behavior at the boundary of $\mathcal{X}$: when x^* is a boundary point, $D(x^*, x)$ could grow as $\Theta(\|x - x^*\|^2)$ in the Euclidean case, as $\Theta(\|x - x^*\|)$ for the negative entropy, or, more generally, as $\Theta(\|x - x^*\|^q)$ for the q-th power regularizer. As a result, when used as a measure of convergence, it is important to rescale the Bregman divergence accordingly in order to avoid inflating – or *deflating* – an algorithm's rate of convergence.

Nonetheless, even if we take this rescaling into account, different instances of (MD) may lead to completely different rates of convergence. Specifically, in terms of absolute values (or

[2]Strictly speaking, the expression we provide for $P_x(y)$ is only valid when $y < x^{q-1}/(1-q)$. The reason for this is that the fractional power regularizer is not strongly convex over $[0,\infty)$, so the corresponding prox-mapping $P_x(y)$ is not well-defined for all values of y. This detail is not important in the calculations that follow, so we disregard it for now.



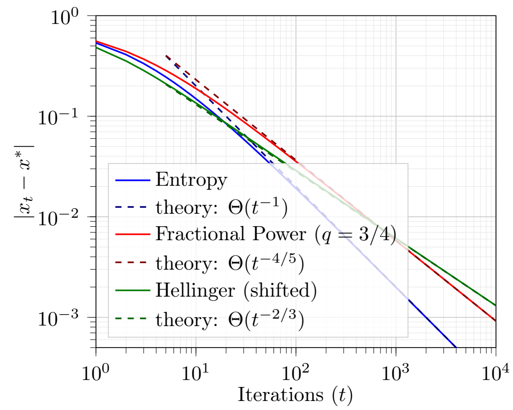


Figure 1: The rate of convergence of (MD) in Examples 3.1–3.4. The Euclidean and shifted Hellinger regularizers lead to a geometric rate of convergence (see left figure); all other examples converge at a polynomial rate, and we provide the theoretically computed rate for comparison (dashed lines in the figure to the right).

norms), we observe a geometric rate in the Euclidean and shifted Hellinger cases, a rate of $\Theta(1/t)$ for the negative entropy, and a rate of $\Theta(1/t^{1/(2-q)})$ for the q-th power regularizer (cf. Fig. 1 above). This is due to the different first-order behavior of the iterative update map $x \leftarrow F(x)$ that underlies (MD), which is itself intimately related to the growth rate of the Bregman divergence near a solution x^* of (VI). We will make this relation precise in the next section.

4. The Legendre exponent and convergence rate analysis

Our goal in this section is to provide a precise link between the geometry induced by a Bregman regularizer near a solution and the convergence rate of the associated Bregman proximal method. The key notion in this regard is that of the *Legendre exponent*, which we define and discuss in detail below.

4.1. The Legendre exponent.

Our starting point is the observation that the strong convexity requirement for h can be equivalently expressed as

$$D(p, x) \geq \tfrac{1}{2}\|p - x\|^2 \quad \text{for all } p \in \mathcal{X},\ x \in \mathcal{X}_h, \tag{22}$$

or, more concisely, $D(p, x) = \Omega(\|p - x\|^2)$ for x near p. Qualitatively, this means that the convergence topology induced by the Bregman divergence of h on $\mathcal{X}$ is *at least as fine* as the ambient norm topology: if a sequence $x_t \in \mathcal{X}_h$, $t = 1, 2, \ldots$, converges to $p \in \mathcal{X}$ in the Bregman sense ($D(p, x_t) \to 0$), then it also converges in the ambient norm topology ($\|x_t - p\| \to 0$). On the other hand, from a quantitative standpoint, the rate of this convergence could be quite different: as we already saw in the previous section, the reverse inequality $D(p, x) = \mathcal{O}(\|p - x\|^2)$ may fail to hold, in which case $\sqrt{D(p, x_t)}$ and $\|x - x_t\|$ would exhibit a different asymptotic behavior.

To quantify this gap, we introduce below the notion of the *Legendre exponent*:

Definition 2. Let h be a Bregman regularizer on $\mathcal{X}$. Then the *Legendre exponent* of h at $p \in \mathcal{X}$ is defined as

$$\beta_h(p) \coloneqq \inf\left\{\beta \in [0,1] : \limsup_{x \to p} \frac{\sqrt{D(p,x)}}{\|x-p\|^{1-\beta}} < \infty\right\} \tag{23}$$

and we will say that h is *tight* at p if the infimum is attained in (23), i.e., if $\beta_h(p)$ is the minimal $\beta \in [0,1]$ such that

$$D(p,x) = \mathcal{O}\big(\|p-x\|^{2(1-\beta)}\big) \quad \text{for } x \text{ near } p. \tag{24}$$

Informally, the Legendre exponent measures the deficit in relative size between ordinary "norm neighborhoods" in $\mathcal{X}$ and the corresponding "Bregman neighborhoods" induced by the sublevel sets of the Bregman divergence. This comparison will play a major role in the sequel, so we proceed with a series of examples and remarks:

Remark 1. A first point to be made is that the restriction $\beta \geq 0$ in Definition 2 is directly imposed by the strong convexity of h. More precisely, given that $D(p,x) = \Omega(\|p-x\|^2)$ for x close to p, we cannot also have $D(p,x) = \mathcal{O}(\|p-x\|^{2+\varepsilon})$ for *any* $\varepsilon > 0$. In particular, this shows that the "norm-like" behavior $D(p,x) = \Theta(\|p-x\|^2)$ corresponds to the case $\beta_h(p) = 0$: any other value of $\beta_h(p)$ would imply a different limiting behavior for $D(p,x)$ as $x \to p$. ♢

Remark 2. At the other end of the spectrum, the exponent $\beta_h(p) = 1$ is representative of the case $\limsup_{x\to p} D(p,x) > 0$. In this case, the ambient norm topology is *strictly coarser* than the Bregman topology: specifically, $D(p,x)$ might remain large even if p and x are topologically close. For example, let $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ be the unit Euclidean ball in $\mathbb{R}^n$ and consider the n-dimensional Hellinger regularizer $h(x) = -\sqrt{1-\|x\|_2^2}$. Then, for all p on the boundary of $\mathcal{X}$ and all $x \in \mathcal{X}_h = \operatorname{int}\mathcal{X}$, we readily get

$$D(p,x) = \frac{1-\langle p,x\rangle}{\sqrt{1-\|x\|_2^2}}. \tag{25}$$

If $n \geq 2$, the limit $\lim_{x\to p} D(p,x)$ may not exist, a fact which has the following counterintuitive consequences: (i) the "Hellinger ball" $\mathcal{B}_r^h(p) \coloneqq \{x \in \mathcal{X}_h : D(p,x) \leq r^2/2\}$ is *not closed* in the Euclidean topology; and (ii) the "Hellinger center" p of $\mathcal{B}_r^h(p)$ actually belongs to the Euclidean boundary of $\mathcal{B}_r^h(p)$. As a result, for all $n \geq 2$, it is straightforward to construct a sequence x_t with $\|x_t - p\|_2 \to 0$, but which remains at *constant* Hellinger divergence relative to p.[3] ♢

Remark 3. For illustration purposes, we compute below the Legendre exponent for each of the running examples of Section 3 (see also forthcoming Table 1):

(1) *Quadratic regularization* (Example 3.1): Since $D(p,x) = (p-x)^2/2$ for all $p,x \in \mathcal{X}$, we have $\beta_h(p) = 0$ for all $p \in \mathcal{X}$.

(2) *Negative entropy* (Example 3.2): For $p = 0$, Eq. (14) gives $D(0,x) = x$, so $\beta_h(0) = 1/2$. Otherwise, for all $p \in \mathcal{X}_h = (0,\infty)$, a Taylor expansion with Lagrange remainder yields $D(p,x) = \mathcal{O}((p-x)^2)$, so $\beta_h(p) = 0$ for all $p \in (0,\infty)$.

[3] For instance, if $n = 2$, the point $x_u = (1-u, \sqrt{2u(1-u)})$ converges to $p = (1,0)$ as $u \to 0^+$, even though $D(p,x_u) = 1$ for all $u \in (0,1)$. Crucially, if $n = 1$, this phenomenon does not occur (cf. Example 3.4 in Section 3).

(3) *Tsallis entropy* (Example 3.3): For $p = 0$, Eq. (17) gives $D(0, x) = x^q/q$, so $\beta_h(0) = \max\{0, 1 - q/2\}$. Otherwise, for all $p \in \mathcal{X}_h = (0, \infty)$, a Taylor expansion yields $D(p, x) = \mathcal{O}((p - x)^2)$, so $\beta_h(p) = 0$ in this case.

(4) *Hellinger regularizer* (Example 3.4): For $p = \pm 1$, Eq. (19) gives $D(\pm 1, x) = \sqrt{(1 \mp x)/(1 \pm x)} = \Theta(|x \mp 1|^{1/2})$, so $\beta_h(\pm 1) = 1 - 1/4 = 3/4$. Instead, if $p \in (-1, 1)$ a Taylor expansion again yields $D(p, x) = \mathcal{O}((p - x)^2)$, so $\beta_h(p) = 0$ in this case. $\lozenge$

Already, a common pattern that emerges from the examples above is that $\beta_h(p) = 0$ whenever p is an interior point. We formalize this intuition below:

Lemma 5. *Suppose that ∇h is locally Lipschitz continuous. Then $\beta_h(p) = 0$ for all $p \in \mathcal{X}_h$; in particular, $\beta_h(p) = 0$ whenever $p \in \operatorname{ri}\mathcal{X}$.*

Proof. Fix some $p \in \mathcal{X}_h$ and suppose that ∇h is locally Lipschitz continuous. Then there exists a neighborhood $\mathcal{U}$ of p in $\mathcal{X}$ and some $\kappa > 0$ such that

$$\|\nabla h(p) - \nabla h(x)\|_* \leq \kappa \|p - x\| \quad \text{for all } x \in \mathcal{U} \cap \mathcal{X}_h. \tag{26}$$

Now, since $\nabla h(p) \in \partial h(p)$, we also have

$$\begin{aligned} D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle &\leq \langle \nabla h(p) - \nabla h(x), p - x \rangle \\ &\leq \|\nabla h(p) - \nabla h(x)\|_* \|p - x\| \leq \kappa \|p - x\|^2 \end{aligned} \tag{27}$$

for all $x \in \mathcal{U} \cap \mathcal{X}_h$. This shows that (24) holds with $\beta = 0$, i.e., $\beta_h(p) = 0$. ■

Remark. Elaborating on (26), we see that the Legendre exponent of h at p is $(1 - \alpha)/2$ whenever ∇h is locally Hölder continuous with exponent α in a neighborhood of p. However, the converse to this statement does not hold – as can be verified immediately from the entropic regularizer. $\lozenge$

4.2. **Convergence rate analysis.** We are now in a position to state our first general result for the convergence rate of (AMP). To do so, we will assume in the rest of this section that x^* is a solution of (VI) satisfying (SOS) and that h is tight at x^* with Legendre exponent $\beta^* := \beta_h(x^*)$. In particular, this means that there exists a neighborhood $\mathcal{U}$ of x^* in $\mathcal{X}$ and a positive constant $\kappa > 0$ such that

$$D(x^*, x) \leq \frac{\kappa}{2} \|x - x^*\|^{2(1 - \beta^*)} \quad \text{for all } x \in \mathcal{U}. \tag{28}$$

We then have the following result.

Theorem 1. *Suppose that Assumptions 1–3 hold and (AMP) is run with a constant step-size $\gamma_t \equiv \gamma$, $t = 1, 2, \dots,$ such that*

$$\gamma \leq \frac{1}{2\varphi L} \quad \text{and} \quad \gamma(1 - a - b)^2 \leq \frac{\mu}{8L^2} \tag{29}$$

where $\varphi = (\sqrt{5} + 1)/2$ is the golden ratio. Then, if x_1 is initialized sufficiently close to x^*, the iterates x_t of (AMP) enjoy the following bounds:

Case 1: *If $\beta^* = 0$, then*

$$D(x^*, x_t) \leq \left(1 - \frac{\mu\gamma}{2\kappa}\right)^{t-1} D(x^*, x_1). \tag{30}$$

	Domain ($\mathcal{X}$)	**Regularizer** (h)	**Legendre Exponent** (β^*)	**Convergence Rate**
Euclidean	arbitrary	$x^2/2$	0	Linear
Entropic	$[0,\infty)$	$x\log x$	$1/2$	$\mathcal{O}(1/t)$
Tsallis	$[0,\infty)$	$[q(1-q)]^{-1}(x-x^q)$	$\max\{0, 1-q/2\}$	$\mathcal{O}(1/t^{q/(2-q)})$
Hellinger	$[-1,1]$	$-\sqrt{1-x^2}$	$3/4$	$\mathcal{O}(1/t^{1/3})$

Table 1: Summary of the Legendre exponents and the associated convergence rates for the 1-dimensional examples of Section 2 ($\mathcal{X} \subseteq \mathbb{R}$). To avoid trivialities, all Legendre exponents refer to boundary points of $\mathcal{X}$.

Case 2: *If* $\beta^* \in (0,1)$*, then*

$$D(x^*, x_t) \leq \frac{D(x^*, x_1)}{(1+\rho\mu\gamma(t-1))^{1/\beta^*-1}} \tag{31}$$

where

$$\rho = \frac{\beta^*}{1-\beta^*} \frac{1}{\max\left(2\kappa^{\frac{1}{1-\beta^*}} D(x^*, x_1)^{-\frac{\beta^*}{1-\beta^*}}, 2^{\frac{\beta^*}{1-\beta^*}}\right)} \tag{32}$$

Before moving on to the proof of Theorem 1, some remarks and corollaries are in order (see also Table 1 for an explicit illustration of the derived rates for Examples 3.1–3.4):

Remark 1. The first point of note is the sharp drop in the convergence rate of (AMP) from geometric, when $\beta^* = 0$, to a power law when $\beta^* > 0$. As we saw in Section 3, this drop is unavoidable, even when $\mathcal{X}$ is 1-dimensional and v is affine; in fact, the calculations of Section 3 show that the guarantees provided by Theorem 1 are, in general, unimprovable. $\diamond$

Remark 2. We should also note that the guarantees of Theorem 1 are all stated in terms of the Bregman divergence, not the ambient norm. Since $D(x^*, x_t) = \Omega(\|x_t - x^*\|^2)$, these bounds can be restated in terms of $\|x_t - x^*\|$, but this conversion is not without loss of information: if the bound $D(x^*, x_t) = \Omega(\|x_t - x^*\|^2)$ is not tight, the actual rate in terms of the norm may be significantly different. This phenomenon was already observed in the 1-dimensional examples of Section 3 where $D(x^*, x_t) = \Theta(\|x_t - x^*\|^{2(1-\beta^*)})$, in which case Theorem 1 gives

$$\|x_t - x^*\| = \mathcal{O}\big(t^{-1/(2\beta^*)}\big) \tag{33}$$

whenever $\beta^* > 0$ (see also Table 1). In general however, the Bregman divergence may grow at different rates along different rays emanating from x^*, so it is not always possible to translate a Bregman-based bound to a norm-based bound (or vice versa). This analysis requires a much closer look at the geometric structure of $\mathcal{X}$, depending on which constraints are active at x^*; we examine this issue at depth in Section 5. $\diamond$

Remark 3. We should also note that, even though Theorem 1 is stated for a constant step-size, our proof allows for a variable step-size γ_t, provided that the step-size conditions (29) continue to hold. In this case, the bounds (30) and (31) respectively become

$$D(x^*, x_t) \leq \prod_{s=1}^{t-1}\left(1 - \frac{\mu\gamma_t}{2\kappa}\right) \cdot D(x^*, x_1) \tag{34}$$

for the Euclidean case ($\beta^* = 0$), and

$$D(x^*, x_t) \leq \frac{D(x^*, x_1)}{(1 + \rho\mu \sum_{s=1}^{t-1} \gamma_s)^{1/\beta^* - 1}} \tag{35}$$

for the Legendre-like case ($0 < \beta^* < 1$). ◊

Remark 4. The constants that appear in in Theorem 1, in particular in the step-size condition (29) are slightly loose for all three prime examples of (AMP). This is the price we pay for the generality of our approach, and can be seen in the proof of Proposition 2 below. ◊

4.3. **Proof of Theorem 1.** We now proceed to the proof of Theorem 1, beginning with a series of helper lemmas and intermediate results tailored to the update structure of (AMP). The first of these lemmas relates the Bregman divergence relative to a base point $p \in \mathcal{X}$ before and after a prox-step.

Lemma 6. *Let $x^+ = P_x(y)$ for some $x \in \mathcal{X}_h$, $y \in \mathcal{Y}$. Then, for all $p \in \mathcal{X}$ and all $v \in \mathrm{PC}(p)$, we have:*

$$D(p, x^+) \leq D(p, x) + \langle y - v, x^+ - p\rangle - D(x^+, x) \tag{36a}$$

$$\leq D(p, x) + \langle y - v, x - p\rangle + \tfrac{1}{2}\|y - v\|_*^2 . \tag{36b}$$

The next lemma extends Lemma 6 to emulate the two-step structure of (AMP):

Lemma 7. *Let $x_i^+ = P_x(y_i)$ for some $x \in \mathcal{X}_h$ and $y_i \in \mathcal{Y}$, $i = 1, 2$. Then, for all $p \in \mathcal{X}$ and all $v \in \mathrm{PC}(p)$, we have:*

$$D(p, x_2^+) \leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 - \tfrac{1}{2}\|x_1^+ - x\|^2. \tag{37}$$

Versions of the above inequalities already exist in the literature, cf. Beck & Teboulle [7], Juditsky et al. [16], Mertikopoulos et al. [23], and references therein. The main novelty in Lemmas 6 and 7 is the extra term involving the polar vector $v \in \mathrm{PC}(p)$; this term plays an important role in the sequel, so we provide complete proofs in Appendix A.

With these preliminaries in hand, we proceed to derive two further inequalities that play a pivotal role in the analysis of (AMP). The first is an immediate – but crucial – corollary of Lemma 7:

Corollary 1. *Let x^* be a solution of (VI). Then, for all $c \geq 0$ and all $t = 1, 2, \dots,$ the iterates of (AMP) satisfy the template inequality*

$$\begin{aligned} D(x^*, x_{t+1}) \leq D(x^*, x_t) &- \gamma_t\langle V_{t+1/2} - cv(x^*), x_{t+1/2} - x^*\rangle \\ &+ \tfrac{1}{2}\gamma_t^2\|V_{t+1/2} - V_t - cv(x^*)\|_*^2 - \tfrac{1}{2}\|x_{t+1/2} - x_t\|^2. \end{aligned} \tag{38}$$

Proof. Since x^* is a solution of (VI), we have $v(x^*) \in -\mathrm{PC}(x^*)$. Eq. (38) then follows by invoking Lemma 7 with $x \leftarrow x_t$, $p \leftarrow x^*$, $v \leftarrow -c\gamma_t v(x^*) \in \mathrm{PC}(x^*)$ and $(y_1, y_2) \leftarrow (-\gamma_t V_t, -\gamma_t V_{t+1/2})$. ■

The second inequality that we derive provides an "energy function" for (AMP), namely

$$E_t = D_t + f_t \tag{39}$$

where

$$D_t = D(x^*, x_t) \quad \text{and} \quad f_t = \gamma_{t-1}^2\|(a + b)V_{t-1/2} - V_{t-1}\|_*^2 \tag{40}$$

with $f_1 = 0$ by convention. The lemma below outlines the Lyapunov properties of E_t and provides much of the heavy lifting for Theorem 1.

Proposition 1. *Suppose that Assumptions 1 and 3 hold and* (AMP) *is run with a step-size such that*

$$\lambda\gamma_t + 4\gamma_t^2 L^2 \leq 1 \quad \textit{for some } \lambda \geq 0 \textit{ and all } t = 1, 2, \dots \tag{41}$$

Then, for all $t = 1, 2, \dots,$ the iterates x_t of (AMP) *satisfy the inequality*

$$\begin{aligned} E_{t+1} \leq E_t - \lambda\gamma_t f_t - \gamma_t\langle v(x_{t+1/2}) - v(x^*), x_{t+1/2} - x^*\rangle \\ - \gamma_t(a+b)\langle v(x^*), x_{t+1/2} - x^*\rangle - \tfrac{1}{2}\|x_{t+1/2} - x_t\|^2 \\ + \gamma_t^2(1-a-b)^2 L^2 \|x_{t+1/2} - x^*\|^2 \\ + 2\gamma_t^2(a+b)^2 L^2\|x_{t+1/2} - x_t\|^2. \end{aligned} \tag{42}$$

Proof. Let $c = 1 - a - b$ so $c \geq 0$ by Assumption 3. Corollary 1 then yields

$$\begin{aligned} D_{t+1} \leq D_t - \gamma_t\langle V_{t+1/2} - v(x^*), x_{t+1/2} - x^*\rangle \\ - \gamma_t(a+b)\langle v(x^*), x_{t+1/2} - x^*\rangle - \tfrac{1}{2}\|x_{t+1/2} - x_t\|^2 \\ + \frac{\gamma_t^2}{2}\|V_{t+1/2} - V_t - cv(x^*)\|_*^2. \end{aligned} \tag{43}$$

Since $V_{t+1/2} = (a+b)V_{t+1/2} + cV_{t+1/2}$, the last term above may be bounded as

$$\begin{aligned} \tfrac{1}{2}\gamma_t^2\|V_{t+1/2} - V_t - cv(x^*)\|_*^2 &\leq \gamma_t^2 c^2\|V_{t+1/2} - v(x^*)\|_*^2 + \gamma_t^2\|(a+b)V_{t+1/2} - V_t\|_*^2 \\ &\leq \gamma_t^2 c^2 L^2 \|x_{t+1/2} - x^*\|_*^2 + \gamma_t^2\|(a+b)V_{t+1/2} - V_t\|_*^2 \\ &= \gamma_t^2(1-a-b)^2 L^2\|x_{t+1/2} - x^*\|^2 + f_{t+1}\,, \end{aligned} \tag{44}$$

where we used Assumption 1 in the second line and the definition (39) of f_t in the last one. Thus, combining Eqs. (43) and (44) and comparing to (42), it suffices to show that

$$2f_{t+1} \leq (1 - \lambda\gamma_t)f_t + 4\gamma_t^2(a+b)^2L^2\|x_{t+1/2} - x_t\|^2 \quad \text{for all } t = 1, 2, \dots \tag{45}$$

We consider two distinct cases for this below.

Case 1: $t = 1$**.** By the definition (39) of f_t and Eqs. (5) and (6), we have:

$$f_2 = \gamma_1^2\|(a+b)V_{3/2} - V_1\|_*^2 = \gamma_1^2(a+b)^2\|v(x_{3/2}) - v(x_1)\|_*^2 \leq \gamma_1^2(a+b)^2L^2\|x_{3/2} - x_1\|^2, \tag{46}$$

where we used the initialization assumption $x_1 = x_{1/2}$ in the second equality and the Lipschitz continuity of v in the last one. Since $f_1 = 0$ by construction, our claim is immediate.

Case 2: $t > 1$**.** By Young's inequality and the Lipschitz continuity of v, we readily obtain

$$\begin{aligned} f_{t+1} &= \gamma_t^2\|(a+b)V_{t+1/2} - V_t\|_*^2 \\ &= \gamma_t^2\big\|(a+b)[v(x_{t+1/2}) - v(x_t)] + b[v(x_t) - v(x_{t-1/2})]\big\|_*^2 \\ &\leq 2\gamma_t^2(a+b)^2\|v(x_{t+1/2}) - v(x_t)\|_*^2 + 2\gamma_t^2 b^2\|v(x_t) - v(x_{t-1/2})\|_*^2 \\ &\leq 2\gamma_t^2(a+b)^2L^2\|x_{t+1/2} - x_t\|^2 + 2\gamma_t^2 b^2 L^2\|x_t - x_{t-1/2}\|^2 \\ &\leq 2\gamma_t^2(a+b)^2L^2\|x_{t+1/2} - x_t\|^2 + 2\gamma_t^2 b^2 L^2 \gamma_{t-1}^2\|V_{t-1/2} - V_{t-1}\|_*^2 \end{aligned} \tag{47}$$

where, in the last line, we used Lemma 3 to bound the difference $x_t - x_{t-1/2}$ as

$$\|x_t - x_{t-1/2}\| = \|P_{x_{t-1}}(-\gamma_{t-1}V_{t-1/2}) - P_{x_{t-1}}(-\gamma_{t-1}V_{t-1})\| \leq \gamma_{t-1}\|V_{t-1/2} - V_{t-1}\|_*. \tag{48}$$

Finally, by Assumption 3, we have $c = 0$ whenever $b > 0$, so $b^2\gamma_{t-1}^2\|V_{t-1/2} - V_{t-1}\|_*^2 = b^2\gamma_{t-1}^2\|(1-c)V_{t-1/2} - V_{t-1}\|_*^2 = b^2 f_t$ for all $t > 1$. Hence, putting everything together, we get

$$f_{t+1} \leq 2\gamma_t^2(a+b)^2L^2\|x_{t+1/2} - x_t\|^2 + 2\gamma_t^2 b^2 L^2 f_t. \tag{49}$$

Eq. (45) then follows by the step-size requirement (41), which implies that $2\gamma_t^2 L^2 \leq (1 - \lambda\gamma_t)/2$. ■

Moving forward, since x^* is a solution of (VI), the first line of (42) yields a negative $\mathcal{O}(\gamma_t)$ contribution to E_t, whereas the third and fourth lines collectively represent a subleading $\mathcal{O}(\gamma_t^2)$ "error term". This decomposition would suffice for the analysis of (AMP) if the coupling term $\langle v(x^*), x_{t+1/2} - x^*\rangle$ did not incur an additional $\mathcal{O}(\gamma_t)$ positive contribution to E_{t+1}. This error term is difficult to control in general, but if x^* satisfies (SOS), it can be mitigated via the following bound:

Lemma 8. *Suppose that Assumption 2 holds. Then, for all $x \in \mathcal{X}$, $x' \in \mathcal{B}$ and all $c \in [0,1]$, we have:*

$$\langle v(x') - cv(x^*), x' - x^*\rangle \geq \tfrac{1}{2}\mu\|x - x^*\|^2 - \mu\|x' - x\|^2. \tag{50}$$

Proof. Since x^* is a solution of (VI) and $c \in [0,1]$, we have $(1-c)\langle v(x^*), x' - x^*\rangle \geq 0$ for all $x' \in \mathcal{X}$. Hence, by Assumption 2, we get

$$\langle v(x') - cv(x^*), x' - x^*\rangle \geq \langle v(x') - v(x^*), x' - x^*\rangle \geq \mu\|x' - x^*\|^2 \tag{51}$$

and our assertion follows from the basic bound $\|x - x^*\|^2 \leq 2\|x - x'\|^2 + 2\|x' - x^*\|^2$. ■

With this ancillary estimate in hand, we may finally sharpen Proposition 1 to obtain a bona fide energy inequality for solutions satisfying (SOS):

Proposition 2. *Suppose that Assumptions 1–3 hold and* (AMP) *is run with a step-size γ_t such that*

$$2\mu\gamma_t + 4\gamma_t^2 L^2 \leq 1 \quad \textit{and} \quad (1-a-b)^2\gamma_t \leq \frac{\mu}{8L^2} \quad \textit{for all } t = 1, 2, \dots \tag{52}$$

Then, for all $t \geq 1$ such that $x_{t+1/2} \in \mathcal{B}$, we have

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{4}\mu\gamma_t\|x_t - x^*\|^2. \tag{53}$$

Proof. Assume that $x_{t+1/2} \in \mathcal{B}$ and set $c = 1 - a - b$. Then, invoking Lemma 8 with $x \leftarrow x_t$ and $x' \leftarrow x_{t+1/2}$, we get

$$\langle v(x_{t+1/2}) - v(x^*), x_{t+1/2} - x^*\rangle + (a+b)\langle v(x^*), x_{t+1/2} - x^*\rangle \geq \tfrac{1}{2}\mu\|x_t - x^*\|^2 - \mu\|x_{t+1/2} - x_t\|^2. \tag{54}$$

Thus, taking $\lambda \leftarrow \mu$ in Proposition 1 (in terms of step-size conditions, the first part of (52) implies (41)) and combining with the above, the bound (42) becomes

$$\begin{aligned} E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{2}\mu\gamma_t\|x_t - x^*\|^2 + \gamma_t^2 c^2 L^2\|x_{t+1/2} - x^*\|^2 \\ - \tfrac{1}{2}\big(1 - 4\gamma_t^2(a+b)^2L^2 - 2\mu\gamma_t\big)\|x_{t+1/2} - x_t\|^2. \end{aligned} \tag{55}$$

Hence, writing $\|x_{t+1/2} - x^*\|^2 \leq 2\|x_{t+1/2} - x_t\|^2 + 2\|x_t - x^*\|^2$ and rearranging, we obtain

$$\begin{aligned} E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{2}\big(\mu\gamma_t - 4\gamma_t^2 c^2 L^2\big)\|x_t - x^*\|^2 \\ - \tfrac{1}{2}\big(1 - 4\gamma_t^2((a+b)^2 + c^2)L^2 - 2\mu\gamma_t\big)\|x_{t+1/2} - x_t\|^2. \end{aligned} \tag{56}$$

Since $a, b, c \geq 0$ and $a + b + c = 1$, we also have $(a+b)^2 + c^2 \leq 1$, so the step-size assumption (52) guarantees that the last term in (56) is nonpositive. Likewise, the second part of (52) gives $\mu\gamma_t - 4\gamma_t^2 c^2 L^2 \geq \frac{1}{2}\mu\gamma_t$, so the energy inequality (53) follows and our proof is complete. ■

Remark. It is worth noting here that applying Lemma 8 in (55) introduces a factor 1/2 for (MD), even though $x_t = x_{t+1/2}$ by construction. Likewise, invoking the second part of the condition (52), which is void for both (MP) and (OMD), leads to a deterioration of the constants in our analysis. ◊

We finally have all the required building blocks in place to prove Theorem 1.

Proof of Theorem 1. Our proof strategy consists of the following basic steps:

(1) We first show that, if the step-size of (AMP) satisfies (29) and x_1 is initialized within a suitable neighborhood of x^*, the sequence of leading states $x_{t+1/2}$, $t = 1, 2, \dots,$ remains within the neighborhood $\mathcal{U} \cap \mathcal{B}$ of x^* where the Legendre bound (28) and (SOS) hold concurrently.

(2) By virtue of this stability result, the energy inequality (53) and the definition of the Legendre exponent allow us to express $D_t = D(x^*, x_t)$ in recursive form as $D_{t+1} \leq D_t - \mathcal{O}\big(D_t^{1/(1-\beta^*)}\big)$ up to an error term that vanishes at a geometric rate. The rates (30) and (31) are then derived by analyzing this recursive inequality for $\beta^* = 0$ and $\beta^* > 0$ respectively.

We now proceed to detail the two steps outlined above.

Step 1: Stability. Take $r > 0$ such that $\mathcal{B}_r^{\mathcal{X}}(x^*) \coloneqq \{x \in \mathcal{X} : \|x - x^*\| \leq r\} \subset \mathcal{U} \cap \mathcal{B}$ and assume further that $x_{1/2} = x_1 \in \mathcal{U} \cap \mathcal{B}$ is such that $D(x^*, x_{1/2}) = D(x^*, x_1) \leq (1 - \lambda) r^2/4$, where $\lambda \in (0, 1)$ is a constant to be determined later. Letting $E_0 \coloneqq D(x^*, x_1)$, we will prove by induction on t that

$$\|x_{t-1/2} - x^*\| \leq r \quad \text{and} \quad E_t \leq E_{t-1}, \tag{57}$$

which will show in particular that $x_{t+1/2} \in \mathcal{U} \cap \mathcal{B}$ for all $t \geq 1$. Indeed:

- For the base case ($t = 1$), we have $\|x_{1/2} - x^*\| \leq \sqrt{2D(x^*, x_{1/2})} \leq r$ and $E_1 = E_0$ by construction, so there is nothing to show.
- For the induction step, assume (57) holds. Then (22) yields

$$\tfrac{1}{2}\|x_t - x^*\|^2 \leq D_t \leq E_t \leq E_1 = D(x^*, x_1) \leq \tfrac{1}{4}(1 - \lambda) r^2 \tag{58}$$

i.e., $x_t \in \mathcal{B}_r^{\mathcal{X}}(x^*)$. Now, to show that $x_{t+1/2} \in \mathcal{B}_r^{\mathcal{X}}(x^*)$, Lemma 6 with $p \leftarrow x^*$, $x \leftarrow x_t$, $y \leftarrow -\gamma_t V_t$ and $v \leftarrow -(a + b)\gamma_t v(x^*)$ gives

$$\begin{aligned} D_{t+1/2} &\leq D_t - \gamma_t \langle V_t - (a + b) v(x^*), x_{t+1/2} - x^* \rangle \\ &\leq D_t - a\gamma_t \langle v(x_t) - v(x^*), x_{t+1/2} - x^* \rangle \\ &\qquad - b\gamma_t \langle v(x_{t-1/2}) - v(x^*), x_{t+1/2} - x^* \rangle \end{aligned} \tag{59}$$

and hence, by Young's inequality and (22), we get

$$\begin{aligned} \tfrac{1}{2}\|x_{t+1/2} - x^*\|^2 &\leq D_t + \gamma_t^2 a \|v(x_t) - v(x^*)\|_*^2 + \gamma_t^2 b \|v(x_{t-1/2}) - v(x^*)\|_*^2 \\ &\qquad + \tfrac{1}{4}(a + b)\|x_{t+1/2} - x^*\|^2 . \end{aligned} \tag{60}$$

Since $a + b \leq 1$, using Assumption 1 and rearranging gives

$$\begin{aligned} \|x_{t+1/2} - x^*\|^2 &\leq 4D_t + 4\gamma_t^2 L^2 \max\{\|x_t - x^*\|^2, \|x_{t-1/2} - x^*\|^2\} \\ &\leq (1 - \lambda) r^2 + 4\gamma_t^2 L^2 r^2 \end{aligned} \tag{61}$$

where we used the fact that $\|x_{t-1/2} - x^*\|^2 \leq r^2$ and $\|x_t - x^*\|^2 \leq 2D_t \leq \frac{1}{2}(1 - \lambda) r^2$ (by the inductive hypothesis and (58) respectively). Therefore, with $2\gamma_t L \leq 1/\varphi < 1$ by assumption, choosing $\lambda = 1/\varphi^2$ gives $\|x_{t+1/2} - x^*\|^2 \leq r^2$, which completes the first part of the induction.

For the second part, our step-size assumption gives

$$2\mu\gamma_t + 4\gamma_t^2 L^2 \leq 2\gamma_t L + 4\gamma_t^2 L^2 \leq 1/\varphi + 1/\varphi^2 = 1. \tag{62}$$

Thus, since $x_{t+1/2} \in \mathcal{B}$, Proposition 2 readily gives

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{4}\mu\gamma_t \|x_t - x^*\|^2 \leq E_t, \tag{63}$$

and the induction is complete.

Step 2: Convergence rate analysis. The local Legendre bound (28) allows us to rewrite (63) as

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \frac{\mu\gamma_t}{2^{1-\alpha}\kappa^{1+\alpha}} D_t^{1+\alpha} \tag{64}$$

where, for convenience, we set $\alpha = \beta^*/(1-\beta^*)$. We now distinguish two cases, depending on whether $\beta^* = 0$ or $\beta^* > 0$.

Case 1: If $\beta^* = 0$, we have $\alpha = 0$ by definition and $\kappa \geq 1$ by (22). Eq. (64) then gives

$$E_{t+1} \leq E_t - \frac{\mu\gamma_t}{2\kappa} D_t - \mu\gamma_t f_t \leq \left(1 - \frac{\mu\gamma_t}{2\kappa}\right) E_t \tag{65}$$

so the bound (30) follows immediately by setting $\gamma_t \equiv \gamma$ for all t.

Case 2: If $\beta^* > 0$, then $\alpha > 0$ too, so we will proceed by rewriting all terms in Eq. (64) in terms of E_t. To that end, we have:

$$\begin{aligned}
E_{t+1} &\leq E_t - \mu\gamma_t f_t - \frac{\mu\gamma_t}{2^{1-\alpha}\kappa^{1+\alpha}} D_t^{1+\alpha} \\
&\leq E_t - \frac{\mu\gamma_t}{D(x^*, x_1)^\alpha} f_t^{1+\alpha} - \frac{\mu\gamma_t}{2^{1-\alpha}\kappa^{1+\alpha}} D_t^{1+\alpha} \\
&\leq E_t - \frac{\mu\gamma_t}{\max(2^{1-\alpha}\kappa^{1+\alpha}, D(x^*, x_1)^\alpha)} \left[D(x^*, x_t)^{1+\alpha} + f_t^{1+\alpha}\right] \\
&\leq E_t - \frac{\mu\gamma_t}{\max(2\kappa^{1+\alpha}, 2^\alpha D(x^*, x_1)^\alpha)} E_t^{1+\alpha}
\end{aligned} \tag{66}$$

where, in the second line, we used (57) to bound f_t as $f_t \leq D(x^*, x_t) + f_t \leq D(x^*, x_1)$, and, in the last line, we used the convexity of $x^{1+\alpha}$ over $\mathbb{R}_+$. The rate (31) then follows by invoking Lemma 6 of [33, p. 46] (recreated in the appendix as Lemma A.1 for ease of reference). ■

5. Finer results for linearly constrained problems

Motivated by applications to game theory and linear programming, our goal in this section will be to take a closer look at the convergence rate of (AMP) for different solution configurations that may arise in practice – and, in particular, in linearly constrained problems. To that end, we begin by revisiting the examples of Section 3.

5.1. Motivating examples, redux.

A common feature of Examples 3.1–3.4 is that the problem's defining vector field vanishes at the solution point under scrutiny. In the series of examples below, we examine the rate of convergence achieved when this is not the case.

Example 5.1 (Euclidean regularization)**.** Consider again the quadratic regularizer of Example 3.1 over $\mathcal{X} = [0, \infty)$, but with $v(x) = x + 1$. The solution of (VI) is still $x^* = 0$ but the update (11) now becomes

$$F(x) = [x - \gamma(x+1)]_+ = [(1-\gamma)x - \gamma]_+ . \tag{67}$$

Since $F(x) = 0$ for all sufficiently small $x > 0$, we readily conclude that x_t converges to x^* in a *finite* number of iterations. ◊

Example 5.2 (Entropic regularization)**.** Under the entropic regularizer of Example 3.2, and taking again $v(x) = x + 1$, the update rule (11) becomes

$$F(x) = x\exp(-\gamma(x+1)) = xe^{-\gamma} + o(x) \sim xe^{-\gamma} \tag{68}$$

i.e., F is a contraction for small $x > 0$. Hence, in contrast to Example 3.2, x_t converges to 0 at a geometric rate, even though the problem's solution lies on the boundary of $\mathcal{X}$. ◊

Example 5.3 (Fractional power)**.** Finally, consider the fractional power regularizer of Example 3.3, again with $v(x) = x + 1$. Then, for $q \in (0, 1)$, the update rule (11) gives

$$F(x) = [x^{q-1} + \gamma(1-q)(x+1)]^{1/(q-1)} = x - \gamma x^{2-q} + o(x^{2-q}) \tag{69}$$

for small $x > 0$. Thus, by Lemma 4, we conclude that x_t converges to 0 as $|x_t - x^*| = \Theta\big(t^{-1/(1-q)}\big)$ and hence $D(x^*, x_t) = \Theta\big(t^{-q/(1-q)}\big)$, which is again faster than the rate predicted by Theorem 1. ◊

Examples 5.1–5.3 already show that the convergence rate of (AMP) can change drastically depending on whether $v(x^*)$ is zero or not. In the example below, we examine in more detail the behavior of the individual coordinates of x_t as a function of the position of $v(x^*)$ relative to $\mathcal{X}$.

Example 5.4 (Higher-dimensional simplices)**.** Consider the canonical two-dimensional simplex $\mathcal{X} = \{(x_1, x_2, x_3) \in \mathbb{R}^3_+ : x_1 + x_2 + x_3 = 1\}$ of $\mathbb{R}^3$ equipped with the entropic regularizer $h(x) = \sum_{i=1}^3 x_i \log x_i$. Consider also the vector field $v(x) = x - p$ with $p = (-\nu_1, -\nu_2, 1)$ for some $\nu_1, \nu_2 \geq 0$, so the solution of (VI) is $x^* = (0, 0, 1)$, an extreme point of $\mathcal{X}$.

Since the Legendre exponent of h at x^* is easily seen to be $\beta_h(x^*) = 1/2$, Theorem 1 would indicate a rate of convergence of $D(x^*, x_t) = \mathcal{O}(1/t)$ or, in terms of norms, $\|x_t - x^*\| = \mathcal{O}(1/t)$. However, this rate can be very pessimistic if, for example, $\nu_1 > 0$. Indeed, in this case, since x_t converges to $x^* = (0, 0, 1)$, the relevant coordinates of $v(x_t)$ will evolve as $v_1(x_t) = x_{1,t} + \nu_1 = \nu_1 + o(1)$ and $v_3(x_t) = x_{3,t} - 1 = o(1)$. Accordingly, since entropic regularization on the simplex leads to the exponential weights update [7]

$$x_{i,t+1} \propto x_{i,t}\exp(-\gamma v_i(x_t)) \quad \text{for all } t \geq 1,\ i = 1, 2, 3, \tag{70}$$

the fact that $\lim_{t\to\infty} x_{3,t} = 1$ readily yields

$$x_{1,t+1} \sim \frac{x_{1,t+1}}{x_{3,t+1}} = \frac{x_{1,t}}{x_{3,t}}\exp(-\gamma v_1(x_t) + \gamma v_3(x_t)) = \frac{x_{1,t}}{x_{3,t}}\exp(-\gamma\nu_1 + o(1)) \tag{71}$$

i.e., $x_{1,t}$ converges to 0 at a *geometric* rate whenever $\nu_1 > 0$.

By symmetry, the argument above yields the same rate for $x_{2,t}$ if $\nu_2 > 0$. However, as we show in Appendix B, if $\nu_2 = 0$, we would have $x_{2,t} = \Theta(1/t)$ no matter the value of ν_1 (and likewise for the rate of $x_{1,t}$ if $\nu_1 = 0$). In other words, the rate provided by Theorem 1 is tight for the coordinate $i \in \{1, 2\}$ with a vanishing drift coefficient ν_i, but not otherwise; we will devote the rest of this section to deriving a formal statement (and proof) of the general principle underlying this observation. ◊

5.2. **Linearly constrained problems.** For concreteness, we will focus in what follows on linearly constrained problems, which is where the structural configurations outlined in the previous examples are more prominent. To simplify the presentation and the analysis, we will identify $\mathcal{V}$ with $\mathbb{R}^n$ endowed with the Euclidean scalar product $\langle\cdot,\cdot\rangle$, and we will not distinguish between primal and dual vectors (meaning in particular that the distinction between normal and polar cones will be likewise blurred).

Formally, we will consider polyhedral domains written in normal form as

$$\mathcal{X} = \{x \in \mathbb{R}^n_+ : Ax = b\} \tag{72}$$

for some matrix $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$.[4] Moreover, we will further assume that $\mathcal{X}$ admits a *Slater point*, i.e., there exists some $x \in \mathcal{X}$ such that $x_i > 0$ for all $i = 1, \dots, n$. This setup is particularly flexible, as it allows us to identify the *active* constraints at $x \in \mathcal{X}$ with the zero components of x.

Elaborating further on this, since $\langle v(x^*), x - x^*\rangle \geq 0$ for all $x \in \mathcal{X}$ and any solution x^* of (VI), we directly infer that $-v(x^*)$ is an element of the normal cone $\mathrm{NC}(x^*)$ to $\mathcal{X}$ at x^*. In our polyhedral setting, $\mathrm{NC}(x^*)$ admits an especially simple representation as

$$\mathrm{NC}(x^*) = \mathrm{row}(A) - \{(\nu_1, \dots, \nu_n) \in \mathbb{R}^n_+ : \nu_i = 0 \text{ whenever } x^*_i = 0\} \tag{73}$$

where $\mathrm{row}(A) = (\ker A)^\perp \subseteq \mathbb{R}^n$ denotes the row space of A [13, Ex. 5.2.6]. As a result, we see that x^* is a solution of (VI) if and only if $v(x^*)$ can be written in the form

$$v(x^*) - \sum_{i\in\mathcal{A}} \nu_i e_i \in \mathrm{row}(A) \tag{74}$$

for an ensemble of non-negative *slackness coefficients* $\nu_i \geq 0$, $i \in \mathcal{A}$, where

$$\mathcal{A} \equiv \mathcal{A}(x^*) = \{i : x^*_i = 0\} \tag{75}$$

denotes the set of inequality constraints of (72) that are active at x^*. In view of all this, we will distinguish the following solution configurations:

Definition 3 (Sharpness)**.** Let $x^* \in \mathcal{X}$ be a solution of (VI) with associated slackness coefficients ν_i, $i \in \mathcal{A}$, as per (74). The set of *sharp* ($\sharp$) and *flat* ($\flat$) directions at x^* are respectively defined as

$$\mathcal{A}_\sharp = \{i \in \mathcal{A} : \nu_i > 0\} \qquad \text{and} \qquad \mathcal{A}_\flat = \{i \in \mathcal{A} : \nu_i = 0\}, \tag{76}$$

and we say that v is *sharp* at x^* if $\mathcal{A}_\sharp = \mathcal{A}$ (or, equivalently, if $\mathcal{A}_\flat = \varnothing$). The *sharpness* of v at x^* is then defined as

$$\nu^* = \min\nolimits_{i\in\mathcal{A}_\sharp} \nu_i. \tag{77}$$

The terminology "sharp" and "flat" alludes to the case where v is a gradient field, and is best illustrated by an example. To wit, let $f(x_1, x_2) = x_1 + \frac{1}{2}(x_2 - 1)^2$ for $x_1, x_2 \geq 0$, so f admits a (unique) global minimizer at $x^* = (0, 1)$. Applying Definition 3 to $v = \nabla f$, we readily get $\mathcal{A}_\sharp = \{1\}$ and $\mathcal{A}_\flat = \{2\}$, reflecting the fact that $f(x_1, 1)$ exhibits a sharp minimum at 0 along x_1 whereas the landscape of $f(0, x_2)$ is flat to first-order around 1 along x_2.

5.3. **Convergence rate analysis.** We are now in a position to state and prove our refinement of Theorem 1 for linearly constrained problems. To that end, following Alvarez et al. [1], we will assume in the rest of this section that (AMP) is run with a Bregman regularizer h that is adapted to the polyhedral structure of $\mathcal{X}$ as per the definition below:

[4] Inequality constraints of the form $Ax \leq b$ can also be accommodated in (72) by introducing the associated slack variables $s = b - Ax \geq 0$. Even though this leads to a more verbose presentation of $\mathcal{X}$, the form (72) is much more convenient in terms of notational overhead, so we will stick with the equality formulation throughout.

Definition 4. Let $\mathcal{X}$ be a polyhedral domain of the general form (72), and let $\theta\colon \mathbb{R}_+ \to \mathbb{R}$ be a continuous function such that *a)* $\theta''(x)$ exists and is positive for all $x > 0$; and *b)* θ'' is locally Lipschitz on $(0, \infty)$. Then, a Bregman regularizer h on $\mathcal{X}$ is said to be *decomposable with kernel* θ if

$$h(x) = \sum_{i=1}^{n} \theta(x_i) \quad \text{for all } x \in \mathcal{X}. \tag{78}$$

In addition to facilitating calculations, the notion of decomposability will further allow us to describe the convergence rate of the iterates of (AMP) near the boundary of $\mathcal{X}$ in finer detail. In fact, as it turns out, the speed of convergence along a given direction will actually be determined by the behavior of the derivative of the Bregman kernel θ near 0.

In this regard, there are two distinct regimes to consider. First, if $\lim_{x\to 0^+} \theta'(x) = -\infty$, it is straightforward to see that $\operatorname{dom}\partial h = \operatorname{ri}\mathcal{X}$ so, by Lemma 1, the iterates x_t of (AMP) will remain in $\operatorname{ri}\mathcal{X}$ for all t; in this case h is essentially smooth – or *Legendre* – in the sense of Rockafellar [36, Chap. 26], and we will refer to it as *steep*. Otherwise, if $\theta'(0)$ exists and is finite, x_t may reach the boundary of $\mathcal{X}$ in a finite number of iterations; we will refer to this case as *non-steep*. The key difference between these two regimes is that, in the non-steep case, the algorithm may achieve convergence in a finite number of steps (at least along certain directions). On the other hand, even though finite-time convergence is not possible in the steep regime, the algorithm's rate of convergence may still depend on the boundary behavior of θ. To illustrate this, we will consider the following concrete cases:

Assumption 4. Let $\theta\colon \mathbb{R}_+ \to \mathbb{R}$ be a kernel function as per Definition 4. Then θ' exhibits one of the following behaviors as $x \to 0^+$:

(*a*) *Euclidean-like:* $\liminf_{x\to 0^+} \theta'(x) > -\infty$.
(*b*) *Entropy-like:* $\liminf_{x\to 0^+} [\theta'(x) + \log x] > -\infty$.
(*c*) *Power-like:* $\liminf_{x\to 0^+} x^p \theta'(x) > -\infty$ for some $p \in (0, 1)$.

Remark. Cases (*a*)–(*c*) above respectively mean that $|\theta'(x)|$ grows as $\mathcal{O}(1)$, $\mathcal{O}(|\log x|)$ or $\mathcal{O}(1/x^p)$ as $x \to 0^+$. Clearly, we have (*a*) $\implies$ (*b*) $\implies$ (*c*) so these cases are not exclusive; nonetheless, to avoid overloading the presentation, when we say that (*b*) holds, we will tacitly imply that (*a*) does not also hold at the same time – and likewise for (*c*). ◊

With all this in hand, we proceed to show that, in linearly constrained problems, (AMP) converges along sharp directions at x^* at an accelerated rate relative to Theorem 1: sublinear rates may become linear, and linear rates transform to convergence in finite time.

Theorem 2. *Suppose that* (AMP) *is run in a polyhedral domain with a decomposable regularizer as per Definition 4. Suppose further that Assumptions 1–4 hold, and that the method's step-size and initialization satisfy the requirements of Theorem 1. Then, for all* $i \in \mathcal{A}_\sharp$, *we have:*

(*a*) *Under Assumption 4(a), there exists some* $T \geq 1$ *such that:*

$$x_{i,t} = 0 \quad \textit{for all } t \geq T \tag{79a}$$

(*b*) *Under Assumption 4(b):*

$$x_{i,t} = \mathcal{O}\big(\exp(-\gamma \nu_{\text{eff}} t/2)\big) \tag{79b}$$

(*c*) *Under Assumption 4(c):*

$$x_{i,t} = \mathcal{O}\big((\gamma \nu_{\text{eff}} t/2)^{-1/p}\big) \tag{79c}$$

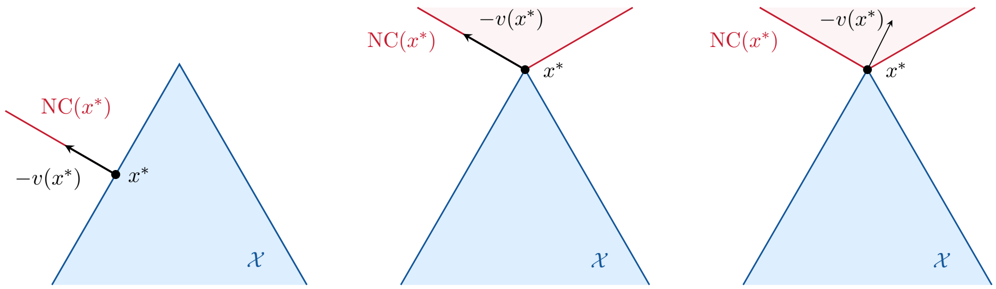


Figure 2: Different solution configurations: a non-extreme solution where v is sharp (left), an extreme solution where v is not sharp (center), and a sharp solution (i.e., an extreme solution where v is sharp; right).

where

$$\nu_{\text{eff}} = \begin{cases} \nu^* & \textit{if } v \textit{ is sharp at } x^* \ (\textit{i.e.}, \mathcal{A}_\flat = \varnothing), \\ \nu^*/\varrho & \textit{otherwise}, \end{cases} \tag{80}$$

and $\varrho \equiv \varrho(A, b, x^*) \geq 1$ *is a positive constant that depends only on* $\mathcal{X}$ *and* x^**.*

Theorem 2 is the main result of this section so, before proving it, some remarks are in order. We begin with the observation that, if the sharp directions at x^* suffice to characterize it, the coordinate-wise guarantees (79) must extend to the full space. This is always so if x^* is an extreme point of $\mathcal{X}$ and v is sharp at x^*, in which case we will say that x^* is itself *sharp*. We then have the following immediate corollary of Theorem 2:

Corollary 2. *If* x^* *is sharp, we have* $\|x_t - x^*\| \to 0$ *at a rate given by* (79a), (79b), *or* (79c), *depending respectively on whether Case (a), (b), or (c) of Assumption 4 holds.*

Proof. First, note that x^* is sharp if and only if $\operatorname{span}\{e_i : i \in \mathcal{A}_\sharp\} + \operatorname{row}(A) = \mathbb{R}^n$: indeed, since $\mathcal{X}$ is a polyhedron, x^* is extreme if and only if $\operatorname{NC}(x^*)$ has nonempty topological interior, and this, combined with (73) and the fact that $\mathcal{A}_\sharp = \mathcal{A}$ (since v is sharp at x^*), proves our assertion. We thus conclude that, for all $i = 1, \dots, n$, there exist $\lambda_{ij} \in \mathbb{R}$, $j \in \mathcal{A}_\sharp$, such that $e_i - \sum_{j \in \mathcal{A}_\sharp} \lambda_{ij} e_j \in \operatorname{row}(A)$, and hence, for all $t = 1, 2, \dots$, we have $x_{i,t} - x_i^* = \sum_{j \in \mathcal{A}_\sharp} \lambda_{ij} x_{j,t}$. Our claim then follows from Theorem 2 and the fact that all norms are equivalent on $\mathbb{R}^n$. ■

We continue with a series of observations elaborating further on Theorem 2.

Remark 1 (Examples). In our series of running examples, the guarantees of Theorem 2 are as follows:

1. *Euclidean regularization* (Example 3.1): With $\theta(x) = x^2/2$, this regularizer satisfies Assumption 4(*a*) because θ' is defined and continuous on all of $\mathbb{R}_+$, so we get convergence along the sharp directions of x^* in a finite number of steps.
2. *Negative entropy* (Example 3.2): The corresponding kernel is $\theta(x) = x \log x$ for $x \geq 0$. Since $\theta'(x) = \log x + 1$ on $(0, \infty)$, θ satisfies Assumption 4(*b*), so the algorithm's rate of convergence along sharp directions is geometric.
3. *Tsallis entropy* (Example 3.3): For $q < 1$, the kernel $\theta(x) = [q(1-q)]^{-1}(x - x^q)$ is differentiable on $(0, \infty)$. Since $\theta'(x) = [q(1-q)]^{-1}(1 - qx^{q-1})$, this kernel satisfies Assumption 4(*c*) with $p = 1 - q$, leading to an $\mathcal{O}(1/t^{1/(1-q)})$ rate of convergence along sharp directions. It is also worth noting here that the Legendre exponent at x^* is upper bounded by p as $\beta^* \leq (1+p)/2$; we defer the details of this calculation to Appendix B.

	Bregman Kernel (θ)	**Generic rate (Theorem 1)**	**Sharp rate (Theorem 2)**
Euclidean	$x^2/2$	Linear	Finite time
Entropic	$x \log x$	$\mathcal{O}(1/t)$	Linear
Tsallis	$[q(1-q)]^{-1}(x-x^q)$	$\mathcal{O}(1/t^{q/(2-q)})$	$\mathcal{O}(1/t^{1/(1-q)})$
Hellinger	$-\sqrt{1-x^2}$	$\mathcal{O}(1/t^{1/3})$	$\mathcal{O}(1/t^2)$

Table 2: Summary of the accelerated rates of convergence observed along sharp directions as a function of the underlying Bregman kernel (cf. Definition 4). The Euclidean, entropic and Tsallis kernels are the prototypical examples of Cases (*a*)–(*c*) of Assumption 4; to avoid trivialities, we only consider the behavior of θ at the boundary of its domain.

(4) *Hellinger regularizer* (Example 3.4): The Hellinger kernel is given by $\theta(x) = -\sqrt{1-x^2}$, so $\theta'(x) = x/(1-x^2)$ for all $x \in (-1,1)$. The behavior of this kernel would then correspond to Assumption 4(*c*) with $p = 1/2$.

To facilitate comparisons with Theorem 1, we juxtapose the corresponding rates in Table 2. $\lozenge$

Remark 2 (Solution configurations). By construction (and the fact that $\mathcal{X}$ admits a Slater point), it is straightforward to verify that v is sharp at x^* if and only if $v(x^*) \in -\operatorname{ri}(\mathrm{NC}(x^*))$; likewise, x^* is itself sharp if and only if $v(x^*) \in -\operatorname{int}(\mathrm{NC}(x^*))$. As we noted in the proof of Corollary 2, the latter condition is equivalent to asking that $\operatorname{span}\{e_i : i \in \mathcal{A}_\sharp\} + \operatorname{row}(A) = \mathbb{R}^n$, a condition which describes precisely the informal requirement that the sharp directions at x^* suffice to characterize it. By contrast, if v is sharp at some non-extreme point x^*, there exists some (nonzero) $z \in \mathrm{TC}(x^*)$ such that $\langle v(x^*), z\rangle = 0$, indicating that the accelerated rates of Theorem 2 cannot be active along the residual direction z. We illustrate these distinct solution configurations in Fig. 2.

$\lozenge$

Remark 3 (Alternative polyhedral representations). Even though every polyhedral domain can be represented in normal form by means of (72) – possibly up to introducing a set of slack variables to account for constraints of the form $Ax \leq b$ – some polyhedra can be represented more succinctly as

$$\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b \text{ and } \langle c_j, x\rangle \in \mathcal{R}_j, j = 1, \dots, m\} \tag{81}$$

for an ensemble of vectors $c_j \in \mathbb{R}^n$ and intervals $\mathcal{R}_j \subseteq \mathbb{R}$ for each of the problem's inequality constraints $j = 1, \dots, m$. At the cost of heavier notation, Theorem 2 can be extended to this setting by considering decomposable Bregman regularizers of the form $h(x) = \sum_{j=1}^m \theta_j(\langle c_j, x\rangle)$, $x \in \mathcal{X}$, where each θ_j is a suitable Bregman kernel on $\mathcal{R}_j$. Mutatis mutandis, the sets of active, sharp and flat constraints can then be defined as in Definition 3, and the guarantees of Theorem 2 would apply to the constraint excess variables $\chi_{j,t} = \langle c_j, x_t\rangle$, $j = 1, \dots, m$. $\lozenge$

Remark 4 (Tightness and the structure of $\mathcal{X}$). It is also important to note that the dependence of ν_{eff} on the structure of $\mathcal{X}$ in the second branch of (80) cannot be lifted. To see this, let

$$\mathcal{X} = \{x \in \mathbb{R}^2_+ : x_1 = \varepsilon x_2\} \tag{82}$$

i.e., $A = [1\ -\varepsilon]$ and $\ker A$ is spanned by the vector $z = (\varepsilon, 1)$. Then, if we take $v(x) = x - p$ with $p_1 \leq 0$ and $p_2 \geq 0$ (so that the origin is a solution), and we equip $\mathcal{X}$ with the Bregman regularizer induced by the entropic kernel $\theta(x) = x \log x$, a straightforward calculation shows that the iterates of (MD) satisfy the recursion

$$\varepsilon\theta'(x_{1,t+1}) + \theta'(x_{2,t+1}) = \big[\varepsilon\theta'(x_{1,t}) + \theta'(x_{2,t})\big] + \gamma\big(\varepsilon x_{1,t} + x_{2,t}\big) - \gamma\langle p, z\rangle. \tag{83}$$

Thus, letting $\chi_t = x_{2,t} = x_{1,t}/\varepsilon$, the above can be rewritten as

$$\varepsilon\theta'(\varepsilon\chi_{t+1}) + \theta'(\chi_{t+1}) = \big[\varepsilon\theta'(\varepsilon\chi_t) + \theta'(\chi_t)\big] + \gamma(\varepsilon^2 + 1)\chi_t - \gamma\langle p, z\rangle \tag{84}$$

and hence, with $\theta'(x) = 1 + \log x$, we finally get

$$\chi_{t+1} = \chi_t \exp\left(-\gamma\frac{\varepsilon^2 + 1}{\varepsilon + 1}\chi_t + \gamma\frac{\langle p, z\rangle}{\varepsilon + 1}\right). \tag{85}$$

Now, if $\langle p, z\rangle < 0$, we readily infer that $\chi_t = x_{2,t}$ converges linearly to 0 at a rate of

$$\chi_t \sim \exp\left(-\frac{|\langle p, z\rangle|}{\varepsilon + 1}\gamma t\right) \tag{86}$$

as predicted by Theorem 2. In particular, if $v(0) = -p = (\nu^*, \nu^*)$ with $\nu^* > 0$, the iterates of (MD) converge geometrically to zero with exponent $\gamma\nu^*$, which matches the estimate of Theorem 2 up to a factor of 1/2 in the exponent. On the contrary, if $v(0) = -p = (\nu^*, 0)$, the term $|\langle -p, z\rangle| = \varepsilon\nu^*$ depends on the linear structure of $\mathcal{X}$ and can be arbitrarily bad as ε goes to zero. This illustrates why one cannot do away with the dependence on the linear structure of $\mathcal{X}$ when v is not sharp at x^*. $\lozenge$

5.4. **Proof of Theorem 2.** We now proceed to the proof of Theorem 2, beginning with two helper lemmas tailored to the polyhedral structure of $\mathcal{X}$. The first is a book-keeping result regarding the subdifferentiability of h.

Lemma 9. *Let h be a decomposable regularizer on $\mathcal{X}$ with kernel θ as per Definition 4. Then the domain of subdifferentiability of h is $\mathcal{X}_h = \{x \in \mathcal{X} : x_i \in \operatorname{dom}\partial\theta$ for all $i = 1, \dots, n\}$ and a continuous selection of ∂h is given by the expression*

$$\nabla h(x) = \sum_{i=1}^{n} \theta'(x_i) e_i \quad \text{for all } x \in \mathcal{X}_h. \tag{87}$$

Proof. See Rockafellar [36, Thm 23.8], whose qualification condition is satisfied thanks to the fact that $\mathcal{X}$ is polyhedral. ■

The second ingredient we will need is a separation result in the spirit of Farkas' lemma.

Lemma 10. *Let $\mathcal{X}$ be a polyhedral domain of the general form (72). Then, for all $x^* \in \mathcal{X}$, there exists $P = P(A, b, x^*) \geq 1$ such that, for all $\mathcal{I} \subseteq \mathcal{A} \equiv \mathcal{A}(x^*)$, at least one of the following holds:*

(a) *$\mathcal{I} \neq \varnothing$ and there exists $i \in \mathcal{A} \setminus \mathcal{I}$ such that $x_i \leq P \max\{x_j : j \in \mathcal{I}\}$ for all $x \in \mathcal{X}$.*
(b) *There exists $z \in \ker A$ such that $\|z\| \leq P$, $z_i = 0$ if $i \in \mathcal{I}$ and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$.*

The proof of Lemma 10 is based on Farkas' lemma so we relegate it to Appendix A and instead proceed to use it to prove our main result for linearly constrained problems.

Proof of Theorem 2. We will consider two main cases, depending on whether $\lim_{x\to 0^+} h'(x) = -\infty$ (the steep case) or $\lim_{x\to 0^+} h'(x) > -\infty$ (the non-steep case). The steep regime will cover Cases (b) and (c) of Assumption 4, whereas the non-steep regime will account for Case (a).

Case 1: the steep regime. We begin by noting that, without loss of generality, Cases (b) and (c) respectively imply that there exist $C \in \mathbb{R}$ and $\delta > 0$ such that, for all $x \in (0, \delta)$, we have:

$$\text{Under Assumption 4}(b)\text{:} \quad \theta'(x) \geq \log x - C \tag{88a}$$

$$\text{Under Assumption 4}(c)\text{:} \quad \theta'(x) \geq -Cx^{-p} \tag{88b}$$

With this in mind, let $r > 0$ be sufficiently small so that the relative neighborhood $\mathcal{B}_r \coloneqq \{x \in \mathcal{X} : \|x - x^*\| \leq r\}$ of x^* in $\mathcal{X}$ satisfies:

- $\mathcal{B}_r \subseteq \mathcal{U} \cap \mathcal{B}$ with $\mathcal{B}$ and $\mathcal{U}$ defined by (SOS) and (28) respectively.
- If $x \in \mathcal{B}_r$ then $x_i < \delta$ for all $i \in \mathcal{A}$.
- If $x \in \mathcal{B}_r$, then $\|v(x) - v(x^*)\|_* \leq \nu^*/(2P)$ with ν^* given by (77) and P given by Lemma 10.

Recall that Step 1 of the proof of Theorem 1 implies that the iterates x_t and half-iterates $x_{t+1/2}$ of (AMP) will remain in $\mathcal{B}_r$ for all t if x_1 is initialized sufficiently close to x^*. Accordingly, with this stability guarantee in hand, we will construct sets $\mathcal{I}_\sharp \subseteq \mathcal{A}_\sharp$, $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$ such that, for all $i \in \mathcal{I} \coloneqq \mathcal{I}_\sharp \cup \mathcal{I}_\flat$,

$$x_{i,t} \leq P^{|\mathcal{I}|} \cdot \begin{cases} \exp(C + P\|\nabla h(x_1)\|_* + PR - \gamma\nu_{\text{eff}}(t-1)/2) & \text{under Assumption 4}(b), \\ C[\gamma\nu_{\text{eff}}(t-1)/2 - P\|\nabla h(x_1)\|_* - PR]_+^{-\frac{1}{p}} & \text{under Assumption 4}(c), \end{cases} \tag{89}$$

where $R \coloneqq \sup_{x \in \mathcal{B}_r} \left\|\sum_{i \notin \mathcal{A}} \theta'(x_i) e_i\right\|_*$, $P \geq 1$ is the constant given by Lemma 10 and ν_{eff} is defined as in (80) with $\varrho = P|\mathcal{A}|$.

We will proceed inductively, starting with $\mathcal{I}_\sharp = \mathcal{I}_\flat = \varnothing$, in which case the stated property holds trivially. For the inductive step, if (89) holds for $\mathcal{I}_\sharp \subsetneq \mathcal{A}_\sharp$ and $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$, we will show that there is some index $j \in \mathcal{A} \setminus \mathcal{I}$ such that (89) still holds for $\mathcal{I} \cup \{j\}$. By iterating this procedure, since the number of active constraints is finite, we will reach $\mathcal{I}_\sharp = \mathcal{A}_\sharp$ and the result of the theorem will follow.

To carry all this out, assume that $\mathcal{I}_\sharp \subsetneq \mathcal{A}_\sharp$, $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$, and apply Lemma 10 to $\mathcal{I} = \mathcal{I}_\sharp \cup \mathcal{I}_\flat$. If the first case of Lemma 10 holds, then $\mathcal{I} \neq \varnothing$ and there exists $i \in \mathcal{A}$ such that

$$x_{i,t} \leq P \max_{j \in \mathcal{I}} x_{j,t} \tag{90}$$

so (89) still holds when i is appended to $\mathcal{I}_\sharp$ or $\mathcal{I}_\flat$ (depending on whether it belongs to $\mathcal{A}_\sharp$ or $\mathcal{A}_\flat$). Otherwise, the second case of Lemma 10 holds and there exists some $z \in \ker A$ with $\|z\| \leq P$, $z_i = 0$ if $i \in \mathcal{I}$, and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$. Since h is steep, this means that $x_{t+1} = P_{x_t}(-\gamma V_{t+1/2})$ belongs to $\mathcal{X}_h = \operatorname{ri} \mathcal{X}$ for all $t = 1, 2, \dots$, so the normal cone $\operatorname{NC}(x_{t+1})$ to $\mathcal{X}$ at x_{t+1} will be the affine hull of $\mathcal{X}$, i.e., $\operatorname{NC}(x_{t+1}) = \operatorname{row}(A)$. Hence, Lemma 1 guarantees that

$$\nabla h(x_{t+1}) - \nabla h(x_t) + \gamma V_{t+1/2} \in \operatorname{row}(A) \tag{91}$$

so, telescoping over $s = 1$ to $t - 1$, we get

$$\nabla h(x_t) - \nabla h(x_1) + \gamma \sum_{s=1}^{t-1} V_{s+1/2} \in \operatorname{row}(A). \tag{92}$$

Taking the scalar product with $z \in \ker A$ then yields

$$\sum_{i=1}^{n} \theta'(x_{i,t}) z_i = \langle \nabla h(x_1), z \rangle - \gamma \sum_{s=1}^{t-1} \langle V_{s+1/2}, z \rangle \tag{93}$$

so, after rearranging and invoking (74) to write $\langle v(x^*), z\rangle = \sum_{i\in\mathcal{A}} \nu_i z_i$, we get

$$\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \theta'(x_{i,t})z_i + \sum_{i\in\mathcal{I}} \theta'(x_{i,t})z_i = \langle \nabla h(x_1), z\rangle - \gamma \sum_{s=1}^{t-1}\sum_{i\in\mathcal{A}} \nu_i z_i + \gamma \sum_{s=1}^{t-1} \langle v(x^*) - V_{s+1/2}, z\rangle - \sum_{i\notin\mathcal{A}} \theta'(x_{i,t})z_i. \tag{94}$$

Now, by the properties we used to construct z, we further have

$$\begin{aligned}\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \theta'(x_{i,t})z_i &\leq P\|\nabla h(x_1)\|_* - \gamma(t-1)\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \nu_i z_i \\ &\quad + \gamma \sum_{s=1}^{t-1} P\|v(x^*) - V_{s+1/2}\|_* + P\Big\|\sum_{i\notin\mathcal{A}} \theta'(x_{i,t})e_i\Big\|_* \\ &\leq P\|\nabla h(x_1)\|_* - \gamma(t-1)\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \nu_i z_i + \tfrac{1}{2}\gamma\nu^*(t-1) + PR. \end{aligned} \tag{95}$$

where the second inequality follows from how we chose $\mathcal{B}_r$ at the beginning of the proof.

We conclude our analysis by distinguishing whether v is sharp at x^* (i.e., if $\mathcal{A}_\flat = \varnothing$ or not).

Case 1: If $\mathcal{A}_\flat = \varnothing$, we have $\mathcal{A}\setminus\mathcal{I} = \mathcal{A}_\sharp\setminus\mathcal{I}_\sharp$ and $\nu_i \geq \nu^*$ for all $i\in\mathcal{A}\setminus\mathcal{I}$, so (95) gives

$$\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \theta'(x_{i,t})z_i + \gamma(t-1)\nu^* z_i \leq P\|\nabla h(x_1)\|_* + \tfrac{1}{2}\gamma\nu^*(t-1) + PR. \tag{96}$$

Choosing the coordinate $j\in\mathcal{A}\setminus\mathcal{I}$ which corresponds to the smallest term in the sum on the left-hand side (LHS), we obtain that

$$(\theta'(x_{j,t}) + \gamma(t-1)\nu^*)(|\mathcal{A}|\setminus\mathcal{I})z_j \leq P\|\nabla h(x_1)\|_* + \tfrac{1}{2}\gamma\nu^*(t-1) + PR \tag{97}$$

and noting that $|\mathcal{A}\setminus\mathcal{I}|z_j \geq 1$ yields

$$\theta'(x_{j,t}) \leq P\|\nabla h(x_1)\|_* - \tfrac{1}{2}\gamma\nu^*(t-1) + PR\,. \tag{98}$$

Case 2: If $\mathcal{A}_\flat \neq \varnothing$, then, since $\mathcal{I}_\sharp \subsetneq \mathcal{A}_\sharp$, the intersection of $\mathcal{A}\setminus\mathcal{I}$ and $\mathcal{A}_\sharp$ is not empty so that, $\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \nu_i z_i \geq \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \nu_i \geq \nu^*$ and the inequality (95) above becomes,

$$\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \theta'(x_{i,t})z_i \leq P\|\nabla h(x_1)\|_* - \tfrac{1}{2}\gamma\nu^*(t-1) + PR\,. \tag{99}$$

Now, choosing j to be the coordinate in $\mathcal{A}\setminus\mathcal{I}$ which minimizes the LHS and bounding $|\mathcal{A}\setminus\mathcal{I}|$ by 1 and $|\mathcal{A}|$, we get that

$$\theta'(x_{j,t})z_j \leq P\|\nabla h(x_1)\|_* - \frac{\gamma\nu^*}{2|\mathcal{A}|}(t-1) + PR\,. \tag{100}$$

Dividing both sides by z_j and using that it lies between 1 and P gives

$$\theta'(x_{j,t}) \leq P\|\nabla h(x_1)\|_* - \frac{\gamma\nu^*}{2P|\mathcal{A}|}(t-1) + PR \tag{101}$$

Therefore, we have shown that, in both cases, there exists $j\in\mathcal{A}\setminus\mathcal{I}$ such that (101) holds, since $P|\mathcal{A}| \geq 1$. Therefore, combining this inequality with (88) we conclude that (89) holds for j, and since $P \geq 1$, we conclude that we can augment $\mathcal{I}_\sharp$ or $\mathcal{I}_\flat$ by j, depending on whether it belongs to $\mathcal{A}_\sharp$ or $\mathcal{A}_\flat$.

Case 2: the non-steep regime. The proof borrows the structure of the first case, though it is more straightforward.

Take $r > 0$ small enough such that $\mathcal{B}_r \coloneqq \{x \in \mathcal{X} : \|x - x^*\| \leq r\}$ satisfies

- $\mathcal{B}_r$ is included in $\mathcal{U} \cap \mathcal{B}$,
- if $x, x' \in \mathcal{B}_r$ then, $\|\nabla h(x) - \nabla h(x')\|_* \leq \frac{\gamma \nu^*}{3P}$ where P is given by Lemma 10. This is possible since ∇h is continuous at x^*.
- if $x \in \mathcal{B}_r$, then $\|v(x) - v(x^*)\|_* \leq \frac{\nu^*}{3P}$.
- No other constraint $x_i = 0$ with $i \notin \mathcal{A}$ becomes active in $\mathcal{B}_r$.

As we have seen in the stability part of the proof of Theorem 1, if (AMP) is started close enough to x^*, then all the iterates x_t and the half-iterates $x_{t+1/2}$ for $s = 1, 2, \ldots$ are contained in $\mathcal{B}_r$.

As above, fix some $t \geq T$. We will build sets $\mathcal{I}_\sharp \subseteq \mathcal{A}_\sharp$, $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$ with the property that that

$$\text{for all } i \in \mathcal{I}_\sharp \cup \mathcal{I}_\flat, x_{i,t} = 0\,. \tag{102}$$

Starting with $\mathcal{I}_\sharp = \mathcal{I}_\flat = \varnothing$, Eq. (102) is trivially verified. Now, take $\mathcal{I}_\sharp \subsetneq \mathcal{A}_\sharp$, $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$ which satisfy the desired property and, as before, apply Lemma 10 with $\mathcal{I} \coloneqq \mathcal{I}_\sharp \cup \mathcal{I}_\flat$. If the first case of Lemma 10 holds, then $\mathcal{I} \neq \varnothing$ and there exists $j \in \mathcal{A}$ such that,

$$x_{j,t} \leq P \max(x_{i,t} : i \in \mathcal{I})\,, \tag{103}$$

which yields the result by adding i to $\mathcal{I}_\sharp$ or $\mathcal{I}_\flat$ depending whether it belongs to $\mathcal{A}_\sharp$ or $\mathcal{A}_\flat$. Otherwise, if the second case holds, there is some $z \in \ker A$ such that $\|z\| \leq P$, $z_i = 0$ if $i \in \mathcal{I}$ and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$. For the sake of contradiction, assume that for all $i \in \mathcal{A} \setminus \mathcal{I}$, $x_{i,t} > 0$. Showing that this results in a contradiction will give us an additional coordinate $j \in \mathcal{A} \setminus \mathcal{I}$ for which $x_{j,t} = 0$ that we will then add to $\mathcal{I}_\sharp$ or $\mathcal{I}_\flat$ as in the first case.

Now, let us determine the normal cone at x_t. Since x_t belongs to $\mathcal{B}_r^{\mathcal{X}}(x^*)$, no other constraint other than the ones corresponding to $\mathcal{I}$ can become active, and these constraints are actually active by the definition of $\mathcal{I}$ and Eq. (102). Hence, the normal cone at x_t (see Eq. (73)) becomes

$$\mathrm{NC}(x_t) = \left\{ -\sum_{i \in \mathcal{I}} \nu_i e_i : (\nu_i)_{i \in \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{I}} \right\} + \mathrm{row}(A)\,, \tag{104}$$

and taking a scalar product between the last inclusion of Lemma 1 and z, we get that,

$$\left\langle \nabla h(x_t) - \nabla h(x_{t-1}) - \gamma V_{t-1/2}, z \right\rangle = 0\,. \tag{105}$$

This means that,

$$\gamma \langle v(x^*), z \rangle = \langle \nabla h(x_t) - \nabla h(x_{t-1}), z \rangle + \gamma \langle v(x^*) - \gamma V_{t-1/2}, z \rangle \leq \frac{2\gamma\nu^*}{3} \tag{106}$$

where the last inequality comes from our definition of $\mathcal{B}_r^{\mathcal{X}}(x^*)$. However, by (74) and the properties of z, we also have

$$\langle v(x^*), z \rangle = \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \nu_i z_i \geq \nu^* \tag{107}$$

which is in contradiction with (106). We may therefore iteratively add coordinates of $\mathcal{A}$ for which $x_{i,t} = 0$, which completes the induction and our proof. ■

6. Concluding remarks

Our results indicate that Euclidean regularization leads to faster trajectory convergence rates near second-order sufficient (SOS) solutions. While this does not contradict the analysis of Nemirovski [27] – which concerns the method's ergodic average in merely monotone problems and advocates the use of entropic regularization in domains with a favorable geometry – it does run contrary to its spirit.

We attribute the source of this discrepancy (at least in the non-sharp case) to the fact that Lipschitz continuity and second-order sufficiency are both norm-based conditions, so it is plausible to expect that norm-based regularizers would lead to better results. This raises the question of what the corresponding rate analysis would give in the case of Bregman-based variants of (LC) and (SOS), e.g., as in the recent works of Bauschke et al. [6], Lu et al. [18], and Antonakopoulos et al. [2]. We defer this analysis to future work.

Acknowledgments

The authors gratefully acknowledge financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the bilateral ANR-NRF grant ALIAS (ANR-19-CE48-0018-01). Part of this work was done while the first author was visiting the Simons Institute for the Theory of Computing.

Appendix A. Auxiliary results

In this appendix, we provide a series of helper lemmas and auxiliary results that we use repeatedly our paper.

Lemmas on numerical sequences. The first two results we provide concern numerical sequences:

Lemma A.1. *Consider two sequences of non-negative real numbers $u_t, \alpha_t \geq 0$, $t = 1, 2, \dots$, such that*

$$u_{t+1} \leq u_t - \alpha_t u_t^{1+r} \quad \text{for some } r > 0 \text{ and all } t = 1, 2, \dots \tag{A.1}$$

Then, for all $t = 1, 2, \dots$, we have:

$$u_{t+1} \leq \frac{u_1}{\left(1 + r u_1^r \sum_{s=1}^t \alpha_s\right)^{1/r}}. \tag{A.2}$$

Proof. See Polyak [33, p. 46]. ■

The second result that we prove here is a slight variant of the above lemma.

Lemma 4. *Suppose that $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ admits the asymptotic expansion*

$$f(x) = x - \lambda x^{1+r} + o(x^{1+r}) \quad \text{as } x \to 0 \tag{16}$$

for positive constants $\lambda, r > 0$. Then, for $u_1 > 0$ small enough, the sequence $u_{t+1} = f(u_t)$, $t = 1, 2, \dots$, converges to 0 at a rate of $u_t \sim (\lambda r t)^{-1/r}$.

Proof. By the assumption on f, there exists some $\varepsilon > 0$ such that

$$x - 2\lambda x^{1+r} \leq f(x) \leq x - \frac{\lambda}{2} x^{1+r} \quad \text{for all } x \in [0, \varepsilon]. \tag{A.3}$$

As a first consequence, if $u_1 \leq \varepsilon$, Lemma A.1 readily implies that u_t converges to zero and that $u_t \leq \varepsilon$ for all t. Moreover, if ε is small enough so that $1 - 2\lambda\varepsilon^r > 0$ and u_1 is positive, this implies that all u_t, for $t = 1, 2, \ldots$, are positive. In particular, it is then valid to consider the sequence u_t^{-r}, $t = 1, 2, \ldots$, for which we get

$$\begin{aligned} u_{t+1}^{-r} - u_t^{-r} &= \left[u_t - \lambda u_t^{1+r} + o\big(u_t^{1+r}\big)\right]^{-r} - u_t^{-r} \\ &= u_t^{-r}(1 - \lambda u_t^r + o\,(u_t^r))^{-r} - u_t^{-r} = r\lambda + o\,(1)\,. \end{aligned} \tag{A.4}$$

Hence, $u_t^{-r} \sim r\lambda t$ which gives the result. ■

Estimates for Bregman proximal steps. The next two results that we provide consider the evolution of the Bregman divergence before and after a prox step (or two):

Lemma 6. *Let $x^+ = P_x(y)$ for some $x \in \mathcal{X}_h$, $y \in \mathcal{Y}$. Then, for all $p \in \mathcal{X}$ and all $v \in \mathrm{PC}(p)$, we have:*

$$D(p, x^+) \le D(p, x) + \langle y - v, x^+ - p\rangle - D(x^+, x) \tag{36a}$$

$$\le D(p, x) + \langle y - v, x - p\rangle + \tfrac{1}{2}\|y - v\|_*^2\,. \tag{36b}$$

Proof. Our proof follows [23, Proposition B.3], but with a slight modification to account for the extra term involving $v \in \mathrm{PC}(p)$. The first step is to invoke the three-point identity (8) to write

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

Then, after rearranging to isolate $D(p, x^+)$, we get

$$\begin{aligned} D(p, x^+) &= D(p, x) - D(x^+, x) - \langle \nabla h(x) - \nabla h(x^+), x^+ - p\rangle \\ &\le D(p, x) - D(x^+, x) - \langle y, x^+ - p\rangle \end{aligned} \tag{A.6}$$

where the inequality in the last line follows from Lemma 1. Hence, given that $\langle v, x^+ - p\rangle \le 0$ by the fact that $v \in \mathrm{PC}(p)$, we readily obtain

$$D(p, x^+) \le D(p, x) - D(x^+, x) - \langle y - v, x^+ - p\rangle\,. \tag{A.7}$$

For the second inequality of the lemma, note that

$$\begin{aligned} -\langle y - v, x^+ - p\rangle &= -\langle y - v, x^+ - x\rangle - \langle y - v, x - x^+\rangle \\ &\le \tfrac{1}{2}\|y - v\|_*^2 + \tfrac{1}{2}\|x^+ - x\|^2 - \langle y - v, x - x^+\rangle \\ &\le \tfrac{1}{2}\|y - v\|_*^2 + D(x^+, x) - \langle y - v, x - x^+\rangle \end{aligned} \tag{A.8}$$

where the penultimate inequality follows directly from Young's inequality and the last one from (22). Our assertion is then obtained by combining this last bound with (A.7). ■

Lemma 7. *Let $x_i^+ = P_x(y_i)$ for some $x \in \mathcal{X}_h$ and $y_i \in \mathcal{Y}$, $i = 1, 2$. Then, for all $p \in \mathcal{X}$ and all $v \in \mathrm{PC}(p)$, we have:*

$$D(p, x_2^+) \le D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 - \tfrac{1}{2}\|x_1^+ - x\|^2. \tag{37}$$

Proof. Our proof follows [23, Proposotion B.4], again with a slight modification to account for the extra terms involving $v \in \mathrm{PC}(p)$. Specifically, applying Lemma 6 with $x_2^+ = P_x(y_2)$ and $v \in \mathrm{PC}(p)$ gives

$$\begin{aligned} D(p, x_2^+) &\le D(p, x) + \langle y_2 - v, x_2^+ - p\rangle - D(x_2^+, x) \\ &\le D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \langle y_2 - v, x_2^+ - x_1^+\rangle - D(x_2^+, x) \end{aligned} \tag{A.9}$$

To lower bound $D(x_2^+, x)$, we invoke again Lemma 6 with $p \leftarrow x_2^+$ and $x_1^+ = P_x(y_1)$; this readily gives

$$D(x_2^+, x_1^+) \le D(x_2^+, x) + \langle y_1, x_1^+ - x_2^+\rangle - D(x_1^+, x) \tag{A.10}$$

and hence, after rearranging to isolate $D(x_2^+, x)$ and substituting the resulting bound in (A.9), we get

$$D(p, x_2^+) \leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \langle y_2 - y_1 - v, x_2^+ - x_1^+\rangle - D(x_2^+, x_1^+) - D(x_1^+, x). \tag{A.11}$$

Thus, by Young's inequality and the strong convexity of h, we finally obtain

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 \\ &\quad + \tfrac{1}{2}\|x_2^+ - x_1^+\|^2 - \tfrac{1}{2}\|x_2^+ - x_1^+\|^2 - \tfrac{1}{2}\|x_1^+ - x\|^2 \\ &\leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 - \tfrac{1}{2}\|x_1^+ - x\|^2 \end{aligned} \tag{A.12}$$

and our proof is complete. ■

A separation result. We now proceed to prove Lemma 10, which we restate below for convenience:

Lemma 10. *Let $\mathcal{X}$ be a polyhedral domain of the general form (72). Then, for all $x^* \in \mathcal{X}$, there exists $P = P(A, b, x^*) \geq 1$ such that, for all $\mathcal{I} \subseteq \mathcal{A} \equiv \mathcal{A}(x^*)$, at least one of the following holds:*

(a) $\mathcal{I} \neq \varnothing$ and there exists $i \in \mathcal{A} \setminus \mathcal{I}$ such that $x_i \leq P \max\{x_j : j \in \mathcal{I}\}$ for all $x \in \mathcal{X}$.

(b) There exists $z \in \ker A$ such that $\|z\| \leq P$, $z_i = 0$ if $i \in \mathcal{I}$ and $P \geq z_i \geq 1$ if $i \in \mathcal{A} \setminus \mathcal{I}$.

Proof. Our claim is trivial if $\mathcal{I} = \mathcal{A}$, so we will focus exclusively on the case $\mathcal{I} \subsetneq \mathcal{A}$. The stated constant $P = P(A, b, x^*)$ will then be obtained as the maximum of 1 and the constants we obtain for each possible $\mathcal{I} \subsetneq \mathcal{A}$.

The proof consists in discussing whether there exists $(\lambda_i)_{i \in \mathcal{A}\setminus\mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A}\setminus\mathcal{I}}$ *not all zero* and $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ such that the inclusion

$$\mathcal{X} \subset \left\{ x \in \mathbb{R}^n : \sum_{i \in \mathcal{A}\setminus\mathcal{I}} \lambda_i x_i = \sum_{i \in \mathcal{I}} \mu_i x_i \right\} \tag{A.13}$$

holds. Case (a) corresponds to the situation where such coeffeicients indeed exist while Case (b) holds when this is not possible.

Case (a). Assume that there exists $(\lambda_i)_{i \in \mathcal{A}\setminus\mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A}\setminus\mathcal{I}}$ *not all zero* and $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ such that (A.13) holds. In this case, $\mathcal{I}$ must be non-empty since otherwise $\mathcal{X}$ would be reduced to $\{0\}$ (see the first inclusion), violating the definition (72) of $\mathcal{X}$. In addition, there is some $i \in \mathcal{A} \setminus \mathcal{I}$ such that $\lambda_i > 0$ and thus we have

$$\forall x \in \mathcal{X},\, x_i \leq \frac{\max\left(|\lambda_j| : j \in \mathcal{I}\right)}{\lambda_i} \max(x_j : j \in \mathcal{I}) \tag{A.14}$$

which corresponds to the first case of the lemma.

Case (b). Otherwise, for all $(\lambda_i)_{i \in \mathcal{A}\setminus\mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A}\setminus\mathcal{I}}$ *not all zero* and $(\mu_i)_{i \in \mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$, (A.13) does not hold.

To interpret this situation, we use the fact that $\mathcal{X}$ is of the general polyhedral form (72) so $\operatorname{aff} \mathcal{X} = x^* + \ker A$ and x^* always satisfies $\sum_{i \in \mathcal{A}\setminus\mathcal{I}} \lambda_i x_i^* = \sum_{i \in \mathcal{I}} \mu_i x_i^* = 0$ so that

$$\begin{aligned} \text{Eq. (A.13)} &\iff \operatorname{aff} \mathcal{X} \subset \left\{ x \in \mathbb{R}^n : \sum_{i \in \mathcal{A}\setminus\mathcal{I}} \lambda_i x_i = \sum_{i \in \mathcal{I}} \mu_i x_i \right\} \\ &\iff \ker A \subset \left\{ x \in \mathbb{R}^n : \sum_{i \in \mathcal{A}\setminus\mathcal{I}} \lambda_i x_i = \sum_{i \in \mathcal{I}} \mu_i x_i \right\} \end{aligned}$$

$$\iff \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \lambda_i e_i - \sum_{i\in\mathcal{I}} \mu_i e_i \in \operatorname{row}(A). \tag{A.15}$$

Therefore, the fact that (A.13) does not hold for all $(\lambda_i)_{i\in\mathcal{A}\setminus\mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A}\setminus\mathcal{I}}$ *not all zero* and $(\mu_i)_{i\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$ means that, the system,

$$\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \lambda_i e_i = \sum_{i\in\mathcal{I}} \mu_i e_i + A^\top r\,, \tag{A.16}$$

with variables $(\lambda_i)_{i\in\mathcal{A}\setminus\mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{A}\setminus\mathcal{I}}$not all zero, $(\mu_i)_{i\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}}, r \in \mathbb{R}^m$, has no solution. Hence, by Motzkin's theorem on the alternative (see e.g., [8, Ex. 1.4.2])[5], this means that the system,

$$\begin{cases} z_i > 0 & \text{for } i \in \mathcal{A}\setminus\mathcal{I} \\ z_i = 0 & \text{for } i \in \mathcal{A} \\ Az = 0 \end{cases} \tag{A.17}$$

admits a solution $z \in \mathbb{R}^n$. Rescaling this solution z and setting P to $\max(\|z\|, \|z\|_\infty)$ then gives the second case. ■

Appendix B. Omitted calculations

In this appendix, we provide some computational details that were left out of the main text to streamline our presentation.

Example 3.4 (Hellinger distance, continuing from p. 8)**.** We proceed to compute the Taylor expansion of F near $x^* = -1$ for the shifted operator $v(x) = x + 1$. Indeed, in this case, the fixed point operator F is given by

$$\begin{aligned} F(x) = P_x(-\gamma v(x)) &= P_x(-\gamma(x+1)) \\ &= \frac{x - \gamma(x+1)\sqrt{1-x^2}}{\sqrt{1 - x^2 + (x - \gamma(x+1)\sqrt{1-x^2})^2}} \\ &= \frac{G(x)}{\sqrt{1-x^2+G(x)^2}}, \end{aligned} \tag{B.1}$$

with $G(x) = x - \gamma(x+1)\sqrt{1-x^2}$. Now, the behavior of G near $x^* = -1$ can be approximated as

$$\begin{aligned} G(x) &= x - \gamma(x+1)^{3/2}(1-x)^{1/2} \\ &= -1 + (x+1) - \gamma(x+1)^{3/2}(2-(x+1))^{1/2} \\ &= -1 + (x+1) - \sqrt{2}\gamma(x+1)^{3/2}\left(1 - \tfrac{1}{4}(x+1) + o(x+1)\right) \\ &= -1 + (x+1) - \sqrt{2}\gamma(x+1)^{3/2} + \tfrac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\Big((x+1)^{5/2}\Big). \end{aligned} \tag{B.2}$$

Another Taylor expansion then yields

$$\begin{aligned} G(x)^2 &= \Big(1 - (x+1) + \sqrt{2}\gamma(x+1)^{3/2} - \tfrac{\sqrt{2}\gamma}{4}(x+1)^{5/2} + o\Big((x+1)^{5/2}\Big)\Big)^2 \\ &= 1 - 2(x+1) + 2\sqrt{2}\gamma(x+1)^{3/2} + (x+1)^2 - \tfrac{\sqrt{2}\gamma}{2}(x+1)^{5/2} + o\Big((x+1)^{5/2}\Big) \end{aligned} \tag{B.3}$$

[5]With the notations of [8, Ex. 1.4.2], the lines of the matrix S are made of the e_i for $i \in \mathcal{A}\setminus\mathcal{I}$ and the lines of the matrix N are the e_i for $i \in \mathcal{I}$, $-e_i$ for $i \in \mathcal{I}$, the lines of A and their opposite.

so the denominator of Eq. (B.1) becomes

$$\begin{aligned}\sqrt{1-x^2+G(x)^2} &= \big((x+1)(2-(x+1)+G(x)^2\big)^2 \\ &= \Big(1+2\sqrt{2}\gamma(x+1)^{3/2}-\tfrac{\sqrt{2}\gamma}{2}(x+1)^{5/2}+o\Big((x+1)^{5/2}\Big)\Big)^2 \\ &= 1+\sqrt{2}\gamma(x+1)^{3/2}-\tfrac{\sqrt{2}\gamma}{4}(x+1)^{5/2}+o\Big((x+1)^{5/2}\Big). \end{aligned} \tag{B.4}$$

Thus, plugging this expansion and Eq. (B.2) into Eq. (B.1) gives

$$\begin{aligned}F(x) &= \frac{-1+(x+1)-\sqrt{2}\gamma(x+1)^{3/2}+\frac{\sqrt{2}\gamma}{4}(x+1)^{5/2}+o\big((x+1)^{5/2}\big)}{1+\sqrt{2}\gamma(x+1)^{3/2}-\frac{\sqrt{2}\gamma}{4}(x+1)^{5/2}+o\big((x+1)^{5/2}\big)} \\ &= \Big(-1+(x+1)-\sqrt{2}\gamma(x+1)^{3/2}+\tfrac{\sqrt{2}\gamma}{4}(x+1)^{5/2}+o\Big((x+1)^{5/2}\Big)\Big) \\ &\qquad\times\left(1-\sqrt{2}\gamma(x+1)^{\frac{3}{2}}+\frac{\sqrt{2}\gamma}{4}(x+1)^{\frac{5}{2}}+o\Big((x+1)^{\frac{5}{2}}\Big)\right) \\ &= -1+(x+1)-2\sqrt{2}\gamma(x+1)^{5/2}+o\Big((x+1)^{5/2}\Big), \end{aligned} \tag{B.5}$$

which gives our assertion when $x^* = -1$. $\lozenge$

Example 5.4 (Three-dimensional simplex, continuing from p. 18)**.** We conclude our treatment of the simplex by showing that $x_{2,t} \sim x_{2,t}/x_{3,t} = \Omega(1/t)$ if $\nu_2 = 0$ but $\nu_1 > 0$. To begin with, we have $v_2(x_t) = x_{2,t} = o(1)$ so, arguing as in the first part of the example, we readily get

$$\frac{x_{1,t+1}}{x_{2,t+1}} = \frac{x_{1,t}}{x_{2,t}}\exp(-\gamma\nu_1+o(1)), \tag{B.6}$$

so $x_{1,t}/x_{2,t}$ converges to 0 at a geometric rate. Accordingly, the quantity of interest $x_{3,t}/x_{2,t}$ can be bounded as

$$\begin{aligned}\frac{x_{3,t+1}}{x_{2,t+1}} &= \frac{x_{3,t}}{x_{2,t}}\exp(\gamma v_2(x_t)-\gamma v_3(x_t)) \\ &= \frac{x_{3,t}}{x_{2,t}}\exp(\gamma x_{2,t}-\gamma(x_{3,t}-1)) = \frac{x_{3,t}}{x_{2,t}}\exp(2\gamma x_{2,t}+\gamma x_{1,t}) \\ &\leq \frac{x_{3,t}}{x_{2,t}}\exp\left(2\gamma\frac{x_{2,t}}{x_{3,t}}+\gamma x_{1,t}\right) \end{aligned} \tag{B.7}$$

Now, since both $\frac{x_{2,t}}{x_{3,t}}$ and $x_{1,t}$ go to zero,

$$\begin{aligned}\frac{x_{3,t+1}}{x_{2,t+1}} &\leq \frac{x_{3,t}}{x_{2,t}}\left(1+2\gamma\frac{x_{3,t}}{x_{2,t}}+\gamma x_{1,t}+o\left(2\frac{x_{3,t}}{x_{2,t}}+x_{1,t}\right)\right) \\ &= \frac{x_{3,t}}{x_{2,t}}+2\gamma+\frac{\gamma x_{1,t}x_{3,t}}{x_{2,t}}+o\left(2+\frac{\gamma x_{1,t}x_{3,t}}{x_{2,t}}\right) \\ &= \frac{x_{3,t}}{x_{2,t}}+2\gamma+o(1). \end{aligned} \tag{B.8}$$

since $x_{1,t}/x_{2,t}$ vanishes as $t \to \infty$. Hence, after telescoping, we conclude that

$$\frac{x_{3,t}}{x_{2,t}} \leq 2\gamma t + o(t), \tag{B.9}$$

which in turn shows that $x_{2,t} \sim x_{2,t}/x_{3,t} = \Omega(1/t)$, as claimed. $\lozenge$

Remark 1 (Connection to the Bregman exponent, continuing from p. 21). We proceed to show here that, in Case (c) of Assumption 4, the Bregman exponent at the solution satisfies $\beta^* \leq (1+p)/2$. Indeed, under our stated assumptions, θ is actually differentiable throughout $(0, \infty)$ so, whenever $0 \leq x' \leq x$, we have

$$\theta(x') - \theta(x) = -\int_{x'}^{x} \theta'(u)\, du. \tag{B.10}$$

Moreover, thanks to Assumption 4(c) with $p < 1$, the integral, $\int_0^x \theta'$is well-defined. Then, letting $x' \to 0^+$ in (B.10), we get $\theta(0) - \theta(x) \leq -\int_0^x \theta'(u)\, du$. On the other hand, Assumption 4(c) implies that both $-\int_0^x \theta'(u)\, du$ and $\theta'(x)(0-x)$ are bounded by $\mathcal{O}(x^{1-p})$ so $\theta(0) - \theta(x) - \theta'(x)(0-x) = \mathcal{O}(x^{1-p})$. Since $\mathcal{A} \neq \varnothing$ and θ' is locally Lipschitz continuous, a similar argument as in the proof of Lemma 5 ultimately yields

$$D(x^*, x) = \sum_{i \in \mathcal{A}} \mathcal{O}\big(x_i^{1-p}\big) + \sum_{i \notin \mathcal{A}} \mathcal{O}\big((x_i^* - x_i)^2\big) = \mathcal{O}\big(\|x^* - x\|^{1-p}\big) \tag{B.11}$$

which shows that the Legendre exponent of h at x^* is at most $(1+p)/2$, as claimed. $\diamond$

References

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

[2] Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. An adaptive mirror-prox algorithm for variational inequalities with singular operators. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[3] Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[4] Auslender, A. and Teboulle, M. Interior projection-like methods for monotone variational inequalities. *Mathematical Programming*, 104:39–68, 2005.

[5] Azizian, W., Iutzeler, F., Malick, J., and Mertikopoulos, P. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *COLT '21: Proceedings of the 34th Annual Conference on Learning Theory*, 2021.

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

[7] Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[8] Ben-Tal, A. and Nemirovski, A. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[9] Bruck Jr., R. E. 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] Chiang, C.-K., Yang, T., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., and Zhu, S. Online optimization with gradual variations. In *COLT '12: Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

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

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

[13] Hiriart-Urruty, J.-B. and Lemaréchal, C. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.

[14] Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 6936–6946, 2019.

[15] Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

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

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

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

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

[20] Martinet, B. Régularisation d'inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis*, 4(R3):154–158, 1970.

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

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

[23] Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. 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.

[24] Minty, G. J. On the generalization of a direct method of the calculus of variations. *Bulletin of the American Mathematical Society*, 73(3):315–321, May 1967.

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

[26] Nemirovski, A. S. Information-based complexity of linear operator equations. *Journal of Complexity*, 8 (2):153–175, 1992.

[27] Nemirovski, A. S. 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.

[28] Nemirovski, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization.* Wiley, New York, NY, 1983.

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

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

[31] Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185:1–35, 2021.

[32] Passty, G. B. 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.

[33] Polyak, B. T. *Introduction to Optimization.* Optimization Software, New York, NY, USA, 1987.

[34] Popov, L. D. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[35] Rakhlin, A. and Sridharan, K. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013.

[36] Rockafellar, R. T. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[37] Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM Journal on Optimization*, 14(5):877–898, 1976.

[38] Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

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

[40] Teboulle, M. A simplified view of first order methods for optimization. *Mathematical Programming*, 170:67–96, 2018.

[41] Tsallis, C. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52: 479–487, 1988.

[42] Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S. P., and Glynn, P. W. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020.