THE RATE OF CONVERGENCE OF BREGMAN PROXIMAL METHODS: LOCAL GEOMETRY VS. REGULARITY VS. SHARPNESS

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 and its optimistic variants – as a function of the local geometry induced by the prox-mapping defining the method. For generality, we focus on local solutions of constrained, non-monotone variational inequalities, and we show that the convergence rate of a given method depends sharply on its associated *Legendre exponent*, a notion that measures the growth rate of the underlying Bregman function (Euclidean, entropic, or other) near a solution. In particular, we show that boundary solutions exhibit a stark separation of regimes between methods with a zero and non-zero Legendre exponent: the former converge at a linear rate, while the latter converge, in general, sublinearly. This dichotomy becomes even more pronounced in linearly constrained problems where methods with entropic regularization achieve a linear convergence rate along sharp directions, compared to convergence in a finite number of steps under Euclidean regularization.

1. Introduction

Bregman proximal methods (BPMs) have a long and rich history in optimization, going back at least to the introduction of the influential mirror descent algorithm by Nemirovski & Yudin [29]. In plain terms, BPMs are first-order (constrained) optimization algorithms that forego Euclidean projections in favor of a more sophisticated "prox-mapping" that minimizes a certain distance-like functional known as the Bregman divergence [9, 11, 20, 29]. When this Bregman divergence is the Euclidean distance squared, one recovers the standard projection-based methods; other than that, depending on the problem's feasible region, different Bregman setups lead to a diverse collection of algorithms, from exponentially weighted gradient descent in the simplex [3, 7, 29], to matrix multiplicative weights on the positive-semidefinite cone [19, 41], variants of Karmarkar's affine scaling algorithm for linear programs [42], etc.

From an operational standpoint, one of the most appealing features of BPMs is that they achieve almost dimension-free convergence rates in problems with a convex structure and a favorable geometry – such as the L^1 ball, the spectraplex, second-order cones, etc. [7, 10, 32]. This is owed to a delicate interplay between the algorithms' non-Euclidean update scheme and the global geometry of the problem's domain. However, these (almost)

---

[*] 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.
The authors are grateful to J. Bolte for many fruitful discussions.

dimension-free guarantees also come with some strings attached: they do not concern the sequence of iterates generated by the method, but only its time average; as a result, the best guarantee that can be achieved after t iterations is $\mathcal{O}(1/t)$. In terms of oracle complexity, this is sufficient for problems that are not strongly convex / strongly monotone, but if one targets finer, geometric convergence rates, the lag induced by averaging cannot be compensated. And, on the other extreme, if the problem is not convex / monotone to begin with, iterate averaging does not provide any quantifiable benefits whatsoever; in this case, only the *actual* iterates generated by the method can be used as candidate solutions.

Our contributions. In this context, our paper seeks to provide a precise characterization of the last-iterate convergence rate of Bregman proximal methods, as a function of the Bregman divergence defining the method and the local geometry that it induces. To treat this question in as general a setting as possible, we focus throughout on variational inequality (VI) problems of the form

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

where $\mathcal{X}$ is a closed convex subset of a finite-dimensional normed space $\mathcal{V}$, and $g\colon \mathcal{X} \to \mathcal{V}^*$ is a (possibly non-monotone) single-valued operator on $\mathcal{X}$ with values in $\mathcal{V}^*$, the dual of $\mathcal{V}$. This problem is a staple of many areas of mathematical programming, game theory and data science, as it provides a template for "optimization beyond minimization" – i.e., for problems where finding an optimal solution does not necessarily involve minimizing a loss function. In particular, in addition to standard minimization problems – which are recovered when $g = \nabla f$ for some smooth function f – the general formulation (VI) includes saddle-point problems, games, complementarity problems, etc.; for an introduction, see [13] and references therein.

In this broad context, we examine the rate of convergence of a wide class of Bregman proximal methods to local solutions of (VI) that satisfy a second-order sufficient condition. Specifically, the class of algorithms we consider includes as special cases (*i*) the original mirror descent (MD) algorithm of [29]; (*ii*) the mirror-prox (MP) method of Nemirovski [31] – which has the same update structure as the Bregman-based algorithm of [4] and contains as a special case the extra-gradient (EG) algorithm of [21]; (*iii*) the so-called optimistic mirror descent (OMD) method of [36] – itself a Bregman analogue of the modified Arrow-Hurwicz algorithm of [35]; etc.

Our first finding is a crisp characterization of last-iterate convergence rate of BPMs in terms of the local geometry induced by the underlying Bregman function near a given solution of (VI). We make this dependence precise via the notion of the *Legendre exponent*, a regularity measure for Bregman methods due to [5], which can roughly be described as the logarithmic ratio of the volume of a Euclidean ball to that of a Bregman ball of the same radius. 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 $\mathcal{O}(t^{-1})$; 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 [15, 17, 23, 27]. By contrast, the Legendre regime ($\beta > 0$) indicates a significant drop in the algorithm's last-iterate convergence speed, even though ergodic convergence rates [31] and results for bilinear games [43] might suggest otherwise.

Subsequently, motivated by applications to game theory and linear programming, we take a closer look at the convergence rate of BPMs across the constraints that are active at a solution x^* of (VI) depending on the position of $g(x^*)$ relative to said constraints. This analysis reveals that Bregman proximal methods have a particularly fine structure: along *sharp directions* (i.e., constraints along which $g(x^*)$ is strictly inward-pointing), BPMs converge (*i*) at a rate of $\mathcal{O}(1/t^{1/(2\beta-1)})$ if $1/2 < \beta < 1$; (*ii*) at a *geometric rate* if $0 < \beta \leq 1/2$ (e.g., for entropic methods); and (*iii*) in a *finite* number of iterations if $\beta = 0$ (cf. Theorem 2). Thus, even though the estimates of Theorem 1 are in general tight, the actual rate of convergence of a Bregman method along different coordinates / constraints could be starkly different – and, in fact, dramatically faster if the solution under study is itself sharp.

The closest antecedent of our work is the conference paper [5] where the Legendre exponent was introduced to analyze the convergence of OMD in *stochastic* VI problems (without considering sharp directions and/or faster identification rates). The stochastic and deterministic settings are obviously very different, both in the challenges involved as well as the rates obtained, so there is no overlap in our analysis and results. Other than that, we are not aware of any comparable results in the literature concerning the radically different convergence landscape of BPMs along active and inactive constraints.

2. Problem setup and preliminaries

To fix notation and terminology, in the rest of our paper $\mathcal{V}$ will denote an 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 $\mathcal{Y}$.

2.1. Blanket assumptions.

Throughout the sequel, we will make the following assumptions for the defining vector field $g\colon \mathcal{X} \to \mathcal{Y}$ of (VI) and the solution $x^* \in \mathcal{X}$ under study:

Assumption 1 (Lipschitz continuity)**.** The vector field g is *L-Lipschitz continuous*, i.e.,

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

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

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

In general, Assumption 2 guarantees that x^* is the unique solution of (VI) in $\mathcal{B}$; we illustrate this in two special cases of interest:

(1) *Minimization problems:* suppose that $g = \nabla f$ for some Lipschitz smooth objective function f on $\mathcal{X}$. Then, 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 $x \in \mathcal{B}$, i.e., x^* is an isolated local minimizer of f.

(2) *Min-max problems:* suppose that $\mathcal{X}$ factorizes as $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ for suitable factor sets $\mathcal{X}_1$, and $\mathcal{X}_2$, let $\mathcal{L}\colon \mathcal{X} \to \mathbb{R}$ be a smooth function on $\mathcal{X}$, and write $g = (\nabla_{x_1}\mathcal{L}, -\nabla_{x_2}\mathcal{L})$ for the min-max gradient of $\mathcal{L}$ (with respect to $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ respectively). Then, any solution $x^* = (x_1^*, x_2^*)$ of (VI) that satisfies Assumption 2 enjoys the local growth bounds $\mathcal{L}(x_1, x_2^*) - \mathcal{L}(x_1^*, x_2^*) = \Omega(\|x_1 - x_1^*\|^2)$ and $\mathcal{L}(x_1^*, x_2^*) - \mathcal{L}(x_1^*, x_2) = \Omega(\|x_2 - x_2^*\|^2)$, implying in turn that x^* is an isolated, hyperbolic saddle-point of $\mathcal{L}$.

More examples satisfying (SOS) include strict Nash equilibria in finite games [14], deterministic Nash policies in (generic) stochastic games [39], etc. Overall, Assumptions 1 and 2 apply to a very wide range of problems, so we will treat them as blanket assumptions throughout.

2.2. **Bregman proximal methods.** As we discussed in the introduction, the main algorithmic template that we will examine for solving (VI) is a general class of first-order algorithms known as *Bregman proximal methods* (BPMs). The defining ingredient of this class is the notion of *Bregman regularizer*, which is defined 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 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

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2. \tag{2.1}$$

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

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle \qquad \text{for all } x \in \mathcal{X}_h,\ p \in \mathcal{X} \tag{2.2}$$

and the induced *prox-mapping* as

$$P_x(y) = \arg\min\nolimits_{x' \in \mathcal{X}} \{\langle y, x - x' \rangle + D(x', x)\} \qquad \text{for all } x \in \mathcal{X}_h,\ y \in \mathcal{Y}. \tag{2.3}$$

Examples of Bregman regularizers are given in Section 3, where we also take an in-depth look at their properties. For now, given a Bregman regularizer on $\mathcal{X}$, the general class of *Bregman proximal methods* (BPMs) that we will consider is defined via the generic recursion

$$x_{t+1/2} = P_{x_t}(-\gamma_t g_t) \qquad x_{t+1} = P_{x_t}(-\gamma_t g_{t+1/2}) \tag{BPM}$$

where (i) $t = 1, 2, \dots$ denotes the method's iteration counter; (ii) $\gamma_t > 0$ is a (non-increasing) step-size sequence; (iii) g_t and $g_{t+1/2}$ are sequences of "oracle signals" that we discuss in detail below. In terms of vocabulary, 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, regarding the sequence of oracle signals g_t and $g_{t+1/2}$ defining (BPM), we will assume throughout that

$$g_{t+1/2} = g(x_{t+1/2}) \qquad \text{for all } t = 1, 2, \dots \tag{2.4}$$

i.e., (BPM) generates a new base state x_{t+1} by taking a Bregman proximal step from x_t with oracle input from the leading state $x_{t+1/2}$. By contrast, the leading state itself can be generated in a number of different ways from x_t, depending on the definition of g_t:

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

$$g_t = a g(x_t) + b g(x_{t-1/2}) \tag{2.5}$$

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

For concreteness, we illustrate below three archetypal Bregman methods that serve as the backbone of the above framework:

(1) *Mirror descent:* following [7, 29, 30], the mirror descent algorithm proceeds recursively as $x^+ = P_x(-\gamma g(x))$, so it can be recovered from (BPM) by taking

$$a = 0,\ b = 0 \quad \text{or, equivalently} \quad g_t = 0 \qquad \text{for all } t = 1, 2, \dots \tag{MD}$$

(2) *Mirror-prox:* following [18, 28], the mirror-prox algorithm corresponds to the choice

$$a = 1,\ b = 0 \quad \text{or, equivalently} \quad g_t = g(x_t) \qquad \text{for all } t = 1, 2, \dots \tag{MP}$$

(3) *Optimistic mirror descent:* originally due to [35] (in the Euclidean case) and [12, 36] (for the general case), the optimistic mirror descent algorithm is obtained by setting

$$a = 0,\ b = 1 \quad \text{or, equivalently} \quad g_t = g(x_{t-1/2}) \quad \text{for all } t = 1, 2, \dots \tag{OMD}$$

These three algorithms are the most widely studied Bregman methods in the literature, so we will use them as running examples throughout.

3. Motivating examples

We now proceed to 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 method. For concreteness, we focus on one-dimensional problems where $\mathcal{X}$ is a closed interval of $\mathbb{R}$ and g is the affine vector field

$$g(x) = x - x^*, \quad x \in \mathbb{R}, \tag{3.1}$$

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

$$x_{t+1} = F(x_t) \quad \text{with} \quad F(x) = P_x(-\gamma g(x)), \tag{3.2}$$

and we will examine the convergence speed of x_t 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:

$$\begin{aligned} &a)\ \text{Prox-domain:} && \mathcal{X}_h = \mathcal{X} \\ &b)\ \text{Bregman divergence:} && D(p,x) = (p-x)^2/2 \\ &c)\ \text{Prox-mapping:} && P_x(y) = [x+y]_+ \end{aligned} \tag{3.3}$$

Consider now the case $x^* = 0$, i.e., $g(x) = x$. Then, for $\gamma \in (0,1)$, the update (3.2) becomes

$$F(x) = x - \gamma x = (1-\gamma)x \tag{3.4}$$

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

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

[1] Note that the requirement "$a+b=1$ if $b>0$" is only intended to ease notation and does not lead to a loss in generality: if $b>0$, we can always rescale γ_t by $a+b$ so the condition $a+b=1$ is satisfied automatically.

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, $h'(x) = 1 + \log x$, which gives

$$
\begin{aligned}
&a)\ \text{Prox-domain:} && \mathcal{X}_h = \operatorname{ri} \mathcal{X} = (0, \infty) \\
&b)\ \text{Bregman divergence:} && D(p, x) = p \log(p/x) + x - p \\
&c)\ \text{Prox-mapping:} && P_x(y) = x \exp(y).
\end{aligned}
\tag{3.5}
$$

Now, taking $g(x) = x$ as in the previous example, the update rule (3.2) becomes

$$
F(x) = x \exp(-\gamma x) = x(1 - \gamma x + o(x)) = x - \gamma x^2 + o(x^2) \quad \text{as } x \to 0. \tag{3.6}
$$

In contrast to (3.4), we now have $F(x) \sim x$ instead of $(1 - \gamma)x$, so F is no longer a contraction. Instead, the iterates of (3.6) may be analyzed by means of the following lemma:

Lemma 1. *Suppose that $f : \mathbb{R}_+ \to \mathbb{R}_+$ admits the asymptotic expansion*

$$
f(x) = x - \lambda x^{1+r} + o(x^{1+r}) \quad \text{as } x \to 0 \tag{3.7}
$$

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 $u_t \sim (\lambda r t)^{-1/r}$.

Thanks to this lemma (which we prove in Appendix A), we readily conclude that x_t converges to 0 at a rate of $D(x^*, x_t) = x_t = |x_t - x^*| \sim 1/(\gamma t)$. ♦

Example 3.3 (Fractional power)**.** Take $\mathcal{X} = [0, \infty)$ and $g(x) = x$ 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, 24, 40]. For this choice of regularizer, we have $h'(x) = (1 - q x^{q-1})/[q(1 - q)]$, and a series of direct calculations gives:[2]

$$
\begin{aligned}
&a)\ \text{Prox-domain:} && \mathcal{X}_h = (0, \infty) \text{ if } q \in (0, 1) \text{ and } \mathcal{X}_h = [0, \infty) \text{ if } q > 1 \\
&b)\ \text{Bregman divergence:} && D(p, x) = \frac{x^q - p^q}{q(1 - q)} - x^{q-1} \frac{x - p}{1 - q} \\
&c)\ \text{Prox-mapping:} && P_x(y) = \big[x^{q-1} - (1 - q) y\big]^{\frac{1}{q-1}} \quad \text{for } q \in (0, 1).
\end{aligned}
\tag{3.8}
$$

Now, when applied to $g(x) = x$, the fractional power variant of (3.2) for $q \in (0, 1)$ gives

$$
F(x) = x \left[1 + \gamma(1 - q) x^{2-q}\right]^{1/(q-1)} = x - \gamma x^{3-q} + o(x^{3-q}) \quad \text{as } x \to 0. \tag{3.9}
$$

Hence, by Lemma 1, we conclude that x_t converges to 0 at a rate of

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

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

[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 this regularizer is not strongly convex over $[0, \infty)$, so the 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.

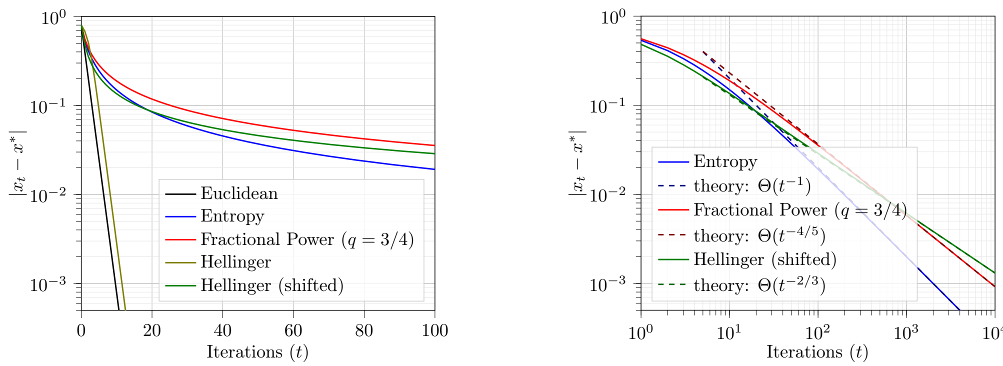

Figure 1: The rate of convergence of (MD) in Examples 3.1–3.4. The Euclidean and shifted Hellinger regularizers lead to a geometric rate (see left figure); all other examples converge at a polynomial rate.

following:

$$
\begin{aligned}
&a)\ \text{Prox-domain:} && \mathcal{X}_h = \operatorname{ri}\mathcal{X} = (-1,1)\\
&b)\ \text{Bregman divergence:} && D(p,x) = \frac{1 - px - \sqrt{(1-p^2)(1-x^2)}}{\sqrt{1-x^2}}\\
&c)\ \text{Prox-mapping:} && P_x(y) = \frac{x + y\sqrt{1-x^2}}{\sqrt{1-x^2+(x+y\sqrt{1-x^2})^2}}.
\end{aligned}
\tag{3.10}
$$

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

$$F(x) = \frac{x - \gamma x\sqrt{1-x^2}}{\sqrt{1-x^2+(x-\gamma x\sqrt{1-x^2})^2}} \sim x - \gamma x \quad \text{as } x \to 0, \tag{3.11}$$

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 $g(x) = x + 1$, a somewhat tedious calculation (which we detail in Appendix B) gives the following Taylor expansion near $x^* = -1$:

$$F(x) = x^* + (x - x^*) - 2\sqrt{2}\gamma(x-x^*)^{5/2} + o\Big((x-x^*)^{5/2}\Big). \tag{3.12}$$

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

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

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 divergence induced by a given regularizer may exhibit a very different behavior at the boundary of $\mathcal{X}$: when x^* is a boundary point, $D(x^*, x)$ grows 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 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, 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

$$D(p,x) \geq \tfrac{1}{2}\|p-x\|^2 \quad \text{for all } p \in \mathcal{X},\ x \in \mathcal{X}_h. \tag{4.1}$$

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 $x_t \in \mathcal{X}_h$, $t = 1, 2, \dots$, 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 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 use the notion of the *Legendre exponent* introduced in [5].

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

$$\beta_h(p) \coloneqq \inf\left\{ \beta \in [0,1] : \limsup_{x \to p} \frac{\sqrt{D(p,x)}}{\|x-p\|^{1-\beta}} < \infty \right\} \tag{4.2}$$

and we say that h is *tight* at p if the infimum is attained in (4.2), i.e., if $\beta_h(p)$ is the minimal $\beta \in [0,1]$ such that

$$D(p,x) = \mathcal{O}\big(\|p-x\|^{2(1-\beta)}\big) \quad \text{for } x \text{ near } p. \tag{4.3}$$

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. Specifically, (*i*) the case $\beta_h(p) = 0$ corresponds to the "norm-like" behavior $D(p,x) = \Theta(\|p-x\|^2)$; (*ii*) any other value $\beta_h(p) \in (0,1)$ indicates a different limiting behavior for $D(p,x)$ as $x \to p$; and, finally, (*iii*) when $\beta_h(p) = 1$ we may have $\limsup_{x \to p} D(p,x) > 0$. In this last case, the ambient norm topology is *strictly coarser* than the Bregman topology in the sense that $D(p, x_t)$ may remain bounded away from zero even if $x_t \to p$ as $t \to \infty$; for a detailed discussion of this phenomenon, cf. [5, Sec. 4] and [33].

For illustration purposes, we compute below the Legendre exponent for each of the running examples of Section 3 (see also 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 $p \in \mathcal{X}$.

	Domain ($\mathcal{X}$)	Regularizer (h)	Legendre Exponent ($\beta_h(p)$)	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 for the 1-dimensional examples of Section 2 at a boundary point p of $\mathcal{X}\subset\mathbb{R}$, and the associated convergence rates.

(2) *Negative entropy* (Example 3.2): For $p=0$, Eq. (3.5) 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) *Tsallis entropy* (Example 3.3): For $p=0$, Eq. (3.8) 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 $p=\pm1$, Eq. (3.10) readily gives $D(\pm1,x)=\sqrt{(1\mp x)/(1\pm x)}=\Theta(|x\mp1|^{1/2})$, so $\beta_h(\pm1)=1-1/4=3/4$. 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.

A common pattern that emerges above is that $\beta_h(p)=0$ whenever p is an interior point. We make this observation precise in Appendix A (Lemma A.5), where we show more generally that $\beta_h(p)=0$ whenever ∇h is (locally) Lipschitz continuous in a neighborhood of p in $\mathcal{X}$.

4.2. **Convergence rate analysis.** We are now in a position to state our first general result for the convergence rate of (BPM). To do so, we will make the blanket assumption 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

$$D(x^*,x)\le\frac{\kappa}{2}\|x-x^*\|^{2(1-\beta^*)}\quad\text{for all } x\in\mathcal{U}. \tag{4.4}$$

We then have the following result.

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

$$\gamma\le\frac{1}{2\varphi L}\quad\text{and}\quad\gamma(1-a-b)^2\le\frac{\mu}{8L^2} \tag{4.5}$$

where $\varphi=(\sqrt{5}+1)/2$ is the golden ratio. If x_1 is initialized sufficiently close to x^*, then the iterates x_t of (BPM) enjoys the bound:

$$D(x^*,x_t)\le D(x^*,x_1)\cdot\begin{cases}\left(1-\frac{\mu\gamma}{2\kappa}\right)^{t-1} & \text{if } \beta^*=0,\\[1ex] [1+\rho\mu\gamma(t-1)]^{1-1/\beta^*} & \text{if } \beta^*\in(0,1),\end{cases} \tag{4.6}$$

where $\rho=\frac{\beta^*}{1-\beta^*}\max\{2\kappa^{\frac{1}{1-\beta^*}}D(x^*,x_1)^{-\frac{\beta^*}{1-\beta^*}},2^{\frac{\beta^*}{1-\beta^*}}\}^{-1}$.

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. A first point of note is the sharp drop in the convergence rate of (BPM) 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 g is affine; in fact, the calculations of Section 3 show that the guarantees of Theorem 1 are, in general, unimprovable. ♦

Remark 2. We note also that the guarantees of Theorem 1 concern 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

$$\|x_t - x^*\| = \mathcal{O}\big(t^{-1/(2\beta^*)}\big) \tag{4.7}$$

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. ♦

Remark 3. 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 (4.5) are satisfied. In this case, the bounds (4.6) becomes

$$D(x^*, x_t) \leq D(x^*, x_1) \begin{cases} \prod_{s=1}^{t-1} \big(1 - \frac{\mu\gamma_s}{2\kappa}\big) & \text{for the Euclidean case } (\beta^* = 0), \\ \big[1 + \rho\mu \sum_{s=1}^{t-1} \gamma_s\big]^{1-1/\beta^*} & \text{for the Legendre case } (0 < \beta^* < 1). \end{cases}$$

The sharpness of these guarantees can be checked as before. ♦

4.3. **Proof of Theorem 1.** We now proceed to the proof of Theorem 1, beginning with a series of intermediate results tailored to the update structure of (BPM). The first of these lemmas relates Bregman divergences before and after a prox-step modulo an element of the polar cone $\mathrm{PC}(p) \coloneqq \{v \in \mathcal{Y} : \langle v, x - p\rangle \leq 0 \text{ for all } x \in \mathcal{X}\}$ of $\mathcal{X}$ at the reference point p.

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

$$D(p, x^+) \leq D(p, x) + \langle y - v, x^+ - p\rangle - D(x^+, x) \tag{4.8a}$$

$$\leq D(p, x) + \langle y - v, x - p\rangle + \tfrac{1}{2}\|y - v\|_*^2 \,. \tag{4.8b}$$

The next lemma extends Lemma 2 to emulate the two-step structure of (BPM):

Lemma 3. *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:*

$$D(p, x_2^+) \leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 - \tfrac{1}{2}\|x_1^+ - x\|^2. \tag{4.9}$$

Versions of the above inequalities already exist in the literature, see e.g., [7, 18, 26]. The novelty in Lemmas 2 and 3 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 (BPM). The first is an immediate corollary of Lemma 3:

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

$$D(x^*, x_{t+1}) \leq D(x^*, x_t) - \gamma_t \langle g_{t+1/2} - cg(x^*), x_{t+1/2} - x^* \rangle \\ + \tfrac{1}{2}\gamma_t^2 \|g_{t+1/2} - g_t - cg(x^*)\|_*^2 - \tfrac{1}{2}\|x_{t+1/2} - x_t\|^2. \tag{4.10}$$

Proof. Since x^* is a solution of (VI), we have $g(x^*) \in -\mathrm{PC}(x^*)$. Eq. (4.10) then follows from Lemma 3 with $x \leftarrow x_t$, $p \leftarrow x^*$, $v \leftarrow -c\gamma_t g(x^*) \in \mathrm{PC}(x^*)$ and $(y_1, y_2) \leftarrow (-\gamma_t g_t, -\gamma_t g_{t+1/2})$. ∎

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

$$E_t = D_t + f_t \tag{4.11}$$

where $D_t = D(x^*, x_t)$ and $f_t = \gamma_{t-1}^2 \|(a+b)g_{t-1/2} - g_{t-1}\|_*^2$ (by convention, we take $f_1 = 0$). The lemma below outlines the Lyapunov properties of E_t.

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

$$\lambda\gamma_t + 4\gamma_t^2 L^2 \leq 1 \quad \textit{for some } \lambda \geq 0 \textit{ and all } t = 1, 2, \dots \tag{4.12}$$

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

$$E_{t+1} \leq E_t - \lambda\gamma_t f_t - \gamma_t \langle g(x_{t+1/2}) - g(x^*), x_{t+1/2} - x^* \rangle \\ - \gamma_t (a+b) \langle g(x^*), x_{t+1/2} - x^* \rangle - \tfrac{1}{2}\|x_{t+1/2} - x_t\|^2 \\ + \gamma_t^2 (1-a-b)^2 L^2 \|x_{t+1/2} - x^*\|^2 \\ + 2\gamma_t^2 (a+b)^2 L^2 \|x_{t+1/2} - x_t\|^2. \tag{4.13}$$

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

$$D_{t+1} \leq D_t - \gamma_t \langle g_{t+1/2} - g(x^*), x_{t+1/2} - x^* \rangle \\ - \gamma_t (a+b) \langle g(x^*), x_{t+1/2} - x^* \rangle - \tfrac{1}{2}\|x_{t+1/2} - x_t\|^2 \\ + \frac{\gamma_t^2}{2} \|g_{t+1/2} - g_t - cg(x^*)\|_*^2. \tag{4.14}$$

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

$$\tfrac{1}{2}\gamma_t^2 \|g_{t+1/2} - g_t - cg(x^*)\|_*^2 \leq \gamma_t^2 c^2 \|g_{t+1/2} - g(x^*)\|_*^2 + \gamma_t^2 \|(a+b)g_{t+1/2} - g_t\|_*^2 \\ \leq \gamma_t^2 c^2 L^2 \|x_{t+1/2} - x^*\|_*^2 + \gamma_t^2 \|(a+b)g_{t+1/2} - g_t\|_*^2 \\ = \gamma_t^2 (1-a-b)^2 L^2 \|x_{t+1/2} - x^*\|^2 + f_{t+1}\,, \tag{4.15}$$

where we used Assumption 1 in the second line and the definition (4.11) of f_t in the last one. Thus, combining Eqs. (4.14) and (4.15) and comparing to (4.13), it suffices to show that

$$2f_{t+1} \leq (1 - \lambda\gamma_t) f_t + 4\gamma_t^2 (a+b)^2 L^2 \|x_{t+1/2} - x_t\|^2 \quad \text{for all } t = 1, 2, \dots \tag{4.16}$$

We consider two distinct cases for this below.

Case 1: $t = 1$**.** By the definition (4.11) of f_t and Eqs. (2.4) and (2.5), we have:

$$\begin{aligned} f_2 &= \gamma_1^2 \|(a+b) g_{3/2} - g_1\|_*^2 \\ &= \gamma_1^2 (a+b)^2 \|g(x_{3/2}) - g(x_1)\|_*^2 \leq \gamma_1^2 (a+b)^2 L^2 \|x_{3/2} - x_1\|^2, \end{aligned} \tag{4.17}$$

where we used the initialization assumption $x_1 = x_{1/2}$ in the second equality and the Lipschitz continuity of g 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 g, we readily obtain

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

where, in the last line, we used Lemma A.4 to bound the difference $x_t - x_{t-1/2}$ as

$$\|x_t - x_{t-1/2}\| = \|P_{x_{t-1}}(-\gamma_{t-1} g_{t-1/2}) - P_{x_{t-1}}(-\gamma_{t-1} g_{t-1})\| \leq \gamma_{t-1} \|g_{t-1/2} - g_{t-1}\|_*. \tag{4.18}$$

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

$$f_{t+1} \leq 2\gamma_t^2 (a+b)^2 L^2 \|x_{t+1/2} - x_t\|^2 + 2\gamma_t^2 b^2 L^2 f_t. \tag{4.19}$$

Eq. (4.16) then follows by the requirement (4.12), 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 (4.13) 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 (BPM) if the coupling term $\langle g(x^*), x_{t+1/2} - x^* \rangle$ did not incur an additional $\mathcal{O}(\gamma_t)$ positive contribution to E_{t+1}. This error term is difficult to control but if x^* satisfies (SOS), we have the following bound.

Lemma 4. *Suppose that Assumption 2 holds. Then, for all $x \in \mathcal{X}$, $x' \in \mathcal{B}$ and all $c \in [0,1]$, we have:*

$$\langle g(x') - c g(x^*), x' - x^* \rangle \geq \tfrac{1}{2} \mu \|x - x^*\|^2 - \mu \|x' - x\|^2. \tag{4.20}$$

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

$$\langle g(x') - c g(x^*), x' - x^* \rangle \geq \langle g(x') - g(x^*), x' - x^* \rangle \geq \mu \|x' - x^*\|^2 \tag{4.21}$$

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* (BPM) *is run with γ_t such that*

$$2\mu\gamma_t + 4\gamma_t^2 L^2 \leq 1 \quad \textit{and} \quad (1-a-b)^2 \gamma_t \leq \frac{\mu}{8L^2} \quad \textit{for all } t = 1, 2, \ldots \tag{4.22}$$

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

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{4}\mu\gamma_t \|x_t - x^*\|^2. \tag{4.23}$$

Proof. Assume that $x_{t+1/2} \in \mathcal{B}$ and set $c = 1 - a - b$. Then, invoking Lemma 4 with $x \leftarrow x_t$ and $x' \leftarrow x_{t+1/2}$, we get

$$\langle g(x_{t+1/2}) - g(x^*), x_{t+1/2} - x^* \rangle + (a+b)\langle g(x^*), x_{t+1/2} - x^* \rangle \geq \tfrac{1}{2}\mu\|x_t - x^*\|^2 - \mu\|x_{t+1/2} - x_t\|^2. \tag{4.24}$$

Thus, taking $\lambda \leftarrow \mu$ in Proposition 1 (in terms of step-size conditions, the first part of (4.22) implies (4.12)) and combining with the above, the bound (4.13) becomes

$$\begin{aligned} E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{2}\mu\gamma_t\|x_t - x^*\|^2 + \gamma_t^2 c^2 L^2 \|x_{t+1/2} - x^*\|^2 \\ - \tfrac{1}{2}\big(1 - 4\gamma_t^2(a+b)^2L^2 - 2\mu\gamma_t\big)\|x_{t+1/2} - x_t\|^2. \end{aligned} \tag{4.25}$$

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

$$\begin{aligned} E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{2}\left(\mu\gamma_t - 4\gamma_t^2c^2L^2\right)\|x_t - x^*\|^2 \\ - \tfrac{1}{2}\big(1 - 4\gamma_t^2((a+b)^2 + c^2)L^2 - 2\mu\gamma_t\big)\|x_{t+1/2} - x_t\|^2. \end{aligned} \tag{4.26}$$

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 (4.22) guarantees that the last term in (4.26) is nonpositive. Likewise, the second part of (4.22) gives $\mu\gamma_t - 4\gamma_t^2c^2L^2 \geq \frac{1}{2}\mu\gamma_t$, so the energy inequality (4.23) follows and our proof is complete. ∎

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 (BPM) satisfies (4.5) and x_1 is initialized sufficiently close to 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 (4.4) and (SOS) both hold.

(2) By virtue of this stability result, the energy inequality (4.23) and the definition of the Legendre exponent allow us to express $D_t = D(x^*, x_t)$ 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 (4.6) 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

$$\|x_{t-1/2} - x^*\| \leq r \quad \text{and} \quad E_t \leq E_{t-1}, \tag{4.27}$$

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 (4.27) holds. Then (4.1) yields

$$\tfrac{1}{2}\|x_t - x^*\|^2 \leq D_t \leq E_t \leq E_1 = D(x^*, x_1) \leq \tfrac{1}{4}(1-\lambda)r^2 \tag{4.28}$$

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 2 with $p \leftarrow x^*$, $x \leftarrow x_t$, $y \leftarrow -\gamma_t g_t$ and $v \leftarrow -(a+b)\gamma_t g(x^*)$ gives

$$\begin{aligned} D_{t+1/2} &\leq D_t - \gamma_t \langle g_t - (a+b) g(x^*), x_{t+1/2} - x^* \rangle \\ &\leq D_t - a\gamma_t \langle g(x_t) - g(x^*), x_{t+1/2} - x^* \rangle \\ &\qquad - b\gamma_t \langle g(x_{t-1/2}) - g(x^*), x_{t+1/2} - x^* \rangle \end{aligned} \tag{4.29}$$

and hence, by Young's inequality and (4.1), we get

$$\begin{aligned} \tfrac{1}{2}\|x_{t+1/2} - x^*\|^2 &\leq D_t + \gamma_t^2 a \|g(x_t) - g(x^*)\|_*^2 + \gamma_t^2 b \|g(x_{t-1/2}) - g(x^*)\|_*^2 \\ &\qquad + \tfrac{1}{4}(a+b)\|x_{t+1/2} - x^*\|^2 . \end{aligned} \tag{4.30}$$

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

$$\begin{aligned} \|x_{t+1/2} - x^*\|^2 &\leq 4D_t + 4\gamma_t^2 L^2 \max\{\|x_t - x^*\|^2, \|x_{t-1/2} - x^*\|^2\} \\ &\leq (1-\lambda) r^2 + 4\gamma_t^2 L^2 r^2 \end{aligned} \tag{4.31}$$

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 (4.28) respectively). Thus, 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. Finally, for the second part, our step-size assumption gives

$$2\mu\gamma_t + 4\gamma_t^2 L^2 \leq 2\gamma_t L + 4\gamma_t^2 L^2 \leq 1/\varphi + 1/\varphi^2 = 1. \tag{4.32}$$

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

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \tfrac{1}{4}\mu\gamma_t \|x_t - x^*\|^2 \leq E_t, \tag{4.33}$$

and the induction is complete.

Step 2: Convergence rate analysis. From (4.33) and the Legendre bound (4.4), we get

$$E_{t+1} \leq E_t - \mu\gamma_t f_t - \frac{\mu\gamma_t}{2^{1-\alpha}\kappa^{1+\alpha}} D_t^{1+\alpha} \qquad \text{with } \alpha = \beta^*/(1-\beta^*). \tag{4.34}$$

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 (4.1). Eq. (4.34) then gives

$$E_{t+1} \leq E_t - \frac{\mu\gamma_t}{2\kappa} D_t - \mu\gamma_t f_t \leq \left(1 - \frac{\mu\gamma_t}{2\kappa}\right) E_t \tag{4.35}$$

so the case $\beta = 0$ of (4.6) 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. (4.34) in terms of E_t. To that end, we have:

$$\begin{aligned} E_{t+1} &\leq E_t - \mu\gamma_t f_t - \frac{\mu\gamma_t}{2^{1-\alpha}\kappa^{1+\alpha}} D_t^{1+\alpha} \\ &\leq E_t - \frac{\mu\gamma_t}{D(x^*, x_1)^\alpha} f_t^{1+\alpha} - \frac{\mu\gamma_t}{2^{1-\alpha}\kappa^{1+\alpha}} D_t^{1+\alpha} \\ &\leq E_t - \frac{\mu\gamma_t}{\max(2^{1-\alpha}\kappa^{1+\alpha}, D(x^*, x_1)^\alpha)} \left[D(x^*, x_t)^{1+\alpha} + f_t^{1+\alpha} \right] \\ &\leq E_t - \frac{\mu\gamma_t}{\max(2\kappa^{1+\alpha}, 2^\alpha D(x^*, x_1)^\alpha)} E_t^{1+\alpha} \end{aligned} \tag{4.36}$$

where, in the second line, we used (4.27) to get $f_t \leq D(x^*, x_t) + f_t \leq D(x^*, x_1)$, and, in the last line, we used the convexity of $x^{1+\alpha}$. The case $\beta \in (0,1)$ of

(4.6) then follows from Lemma 6 of [34, p. 46] (recreated in the appendix as Lemma A.1).

This concludes our proof. ■

5. Finer results for linearly constrained problems

Our goal in this last section is to take a closer look at the convergence rate of (BPM) for different solution configurations that arise 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 study. 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 $g(x) = x + 1$. The solution of (VI) is still $x^* = 0$ but the update (3.2) now becomes

$$F(x) = [x - \gamma(x+1)]_+ = [(1-\gamma)x - \gamma]_+ . \tag{5.1}$$

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 $g(x) = x + 1$, the update rule (3.2) becomes

$$F(x) = x \exp(-\gamma(x+1)) = xe^{-\gamma} + o(x) \sim xe^{-\gamma} \tag{5.2}$$

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 $g(x) = x + 1$. Then, for $q \in (0, 1)$, the update rule (3.2) gives

$$F(x) = [x^{q-1} + \gamma(1-q)(x+1)]^{1/(q-1)} = x - \gamma x^{2-q} + o(x^{2-q}) \tag{5.3}$$

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

Examples 5.1–5.3 show that the convergence rate of (BPM) can change drastically depending on whether $g(x^*)$ 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 $g(x^*)$ 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 $g(x) = x - p$ 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 $g(x_t)$

will evolve as $g_1(x_t) = x_{1,t} + \nu_1 = \nu_1 + o(1)$ and $g_3(x_t) = x_{3,t} - 1 = o(1)$. Accordingly, since entropic regularization on the simplex leads to the exponential weights update [7]

$$x_{i,t+1} \propto x_{i,t} \exp(-\gamma g_i(x_t)) \quad \text{for all } t \geq 1,\ i = 1, 2, 3, \tag{5.4}$$

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

$$x_{1,t+1} \sim \frac{x_{1,t+1}}{x_{3,t+1}} = \frac{x_{1,t}}{x_{3,t}} \exp(-\gamma g_1(x_t) + \gamma g_3(x_t)) = \frac{x_{1,t}}{x_{3,t}} \exp(-\gamma \nu_1 + o(1)) \tag{5.5}$$

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 focus below on linearly constrained problems: the structural configurations outlined in the previous examples are there more prominent. To simplify the presentation, we identify $\mathcal{V}$ with $\mathbb{R}^n$ endowed with the Euclidean scalar product $\langle\cdot,\cdot\rangle$, and we do not distinguish between primal and dual vectors (meaning in particular that the distinction between normal and polar cones are likewise blurred).

Formally, we consider polyhedral domains written in normal form as

$$\mathcal{X} = \{x \in \mathbb{R}^n_+ : Ax = b\} \tag{5.6}$$

for some matrix $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$.[3] Moreover, to avoid trivialities, we 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 $\langle g(x^*), x - x^*\rangle \geq 0$ for all $x \in \mathcal{X}$ and any solution x^* of (VI), we directly infer that $-g(x^*)$ 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

$$\mathrm{NC}(x^*) = \mathrm{row}(A) - \{(\nu_1, \dots, \nu_n) \in \mathbb{R}^n_+ : \nu_i = 0 \text{ whenever } x^*_i = 0\} \tag{5.7}$$

where $\mathrm{row}(A) = (\ker A)^\perp \subseteq \mathbb{R}^n$ denotes the row space of A [16, Ex. 5.2.6]. As a result, we see that x^* is a solution of (VI) if and only if $g(x^*)$ can be written in the form

$$g(x^*) - \sum_{i\in\mathcal{A}} \nu_i e_i \in \mathrm{row}(A) \tag{5.8}$$

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

$$\mathcal{A} \equiv \mathcal{A}(x^*) = \{i : x^*_i = 0\} \tag{5.9}$$

denotes the set of inequality constraints of (5.6) that are active at x^*. With all this in mind, we will distinguish the following solution configurations (see also Fig. 2).

[3]Inequality constraints of the form $Ax \leq b$ can also be accommodated in (5.6) 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 (5.6) is much more convenient in terms of notational overhead, so we stick with the equality formulation throughout.

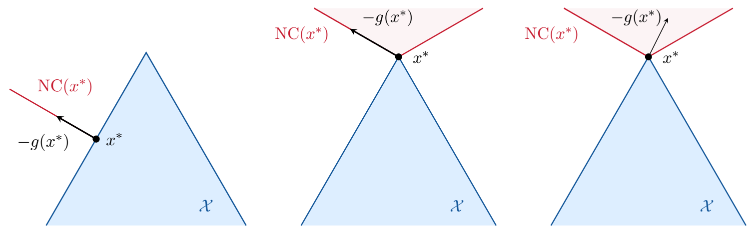

Figure 2: Different solution configurations: a non-extreme solution where g is sharp (left), an extreme solution where g is not sharp (center), and a sharp solution (right).

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 (5.8). The set of *sharp* ($\sharp$) and *flat* ($\flat$) directions at x^* are respectively defined as

$$\mathcal{A}_\sharp = \{i \in \mathcal{A} : \nu_i > 0\} \qquad \text{and} \qquad \mathcal{A}_\flat = \{i \in \mathcal{A} : \nu_i = 0\}, \tag{5.10}$$

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

$$\nu^* = \min_{i \in \mathcal{A}_\sharp} \nu_i \tag{5.11}$$

and, finally, if x^* is an extreme point of $\mathcal{X}$, we will say that x^* is itself *sharp*.

The terminology "sharp" and "flat" alludes to the case where g 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 $g = \nabla f$, 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 [1], we will assume in the rest of this section that (BPM) is run with a Bregman regularizer h that is adapted to the polyhedral structure of $\mathcal{X}$ as per the definition below:

Definition 4. Let $\mathcal{X}$ be a polyhedral domain of the form (5.6). Then, a Bregman regularizer h on $\mathcal{X}$ is said to be *decomposable with kernel* θ if

$$h(x) = \sum_{i=1}^{n} \theta(x_i) \quad \text{for all } x \in \mathcal{X}. \tag{5.12}$$

for some continuous function $\theta \colon \mathbb{R}_+ \to \mathbb{R}$ such that (*i*) $\theta''(z) > 0$ for all $z > 0$; and (*ii*) θ'' is locally Lipschitz on $(0, \infty)$.

In addition to facilitating calculations, the notion of decomposability will further allow us to describe the convergence rate of the iterates of (BPM) 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 A.2 in Appendix A, the iterates x_t of (BPM) will remain in $\operatorname{ri} \mathcal{X}$ for all t; in this case h is essentially smooth – or *Legendre* – in the sense of [37, 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 assume (b), 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 below to show that, in linearly constrained problems, (BPM) 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* (BPM) *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
$$x_{i,t} = 0 \quad \textit{for all } t \geq T \tag{5.13a}$$

(b) Under Assumption 4(b):
$$x_{i,t} = \mathcal{O}\big(\exp(-\gamma\nu_{\text{eff}} t/2)\big) \tag{5.13b}$$

(c) Under Assumption 4(c):
$$x_{i,t} = \mathcal{O}\big((\gamma\nu_{\text{eff}} t/2)^{-1/p}\big) \tag{5.13c}$$

where
$$\nu_{\text{eff}} = \begin{cases} \nu^* & \textit{if } g \textit{ is sharp at } x^* \ (\textit{i.e., } \mathcal{A}_\flat = \varnothing), \\ \nu^*/\varrho & \textit{otherwise,} \end{cases} \tag{5.14}$$

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

In particular, if x^* is sharp, we have the following immediate corollary of Theorem 2.

Corollary 2. *If x^* is sharp, then $\|x_t - x^*\| \to 0$ at a rate given by* (5.13a), (5.13b), *or* (5.13c), *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 (5.7) and the fact that $\mathcal{A}_\sharp = \mathcal{A}$ (since g is sharp at x^*),

	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 depending on the Bregman kernel (cf. Definition 4). The Euclidean, entropic and Tsallis kernels are the prototypical examples of Cases *(a)*–*(c)* of Assumption 4.

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 $\mathbb{R}^n$. ■

To facilitate comparisons with the non-sharp regime, the guarantees of Theorem 2 are juxtaposed with those of Theorem 1 in Table 2. Beyond this comparison, Theorem 2 is the main result of this section so, before proving it, some further remarks are in order.

Remark 1 (Solution configurations). By construction (and the fact that $\mathcal{X}$ admits a Slater point), it is straightforward to verify that g is sharp at x^* if and only if $g(x^*) \in -\operatorname{ri}(\operatorname{NC}(x^*))$; likewise, x^* is itself sharp if and only if $g(x^*) \in -\operatorname{int}(\operatorname{NC}(x^*))$. 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 g is sharp at some non-extreme point x^*, there exists some (nonzero) $z \in \operatorname{TC}(x^*)$ such that $\langle g(x^*), z\rangle = 0$, 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. ♦

Remark 2 (Tightness and the structure of $\mathcal{X}$). We underline that the dependence of ν_{eff} on the structure of $\mathcal{X}$ in the second branch of (5.14) cannot be lifted. To see this, let

$$\mathcal{X} = \{x \in \mathbb{R}^2_+ : x_1 = \varepsilon x_2\} \tag{5.15}$$

i.e., $A = [1 \; -\varepsilon]$ and $\ker A$ is spanned by the vector $z = (\varepsilon, 1)$. Then, if we take $g(x) = x - p$ 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

$$\varepsilon\theta'(x_{1,t+1}) + \theta'(x_{2,t+1}) = \left[\varepsilon\theta'(x_{1,t}) + \theta'(x_{2,t})\right] + \gamma\big(\varepsilon x_{1,t} + x_{2,t}\big) - \gamma\langle p, z\rangle. \tag{5.16}$$

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

$$\varepsilon\theta'(\varepsilon\chi_{t+1}) + \theta'(\chi_{t+1}) = \left[\varepsilon\theta'(\varepsilon\chi_t) + \theta'(\chi_t)\right] + \gamma(\varepsilon^2 + 1)\chi_t - \gamma\langle p, z\rangle \tag{5.17}$$

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

$$\chi_{t+1} = \chi_t \exp\left(-\gamma \frac{\varepsilon^2 + 1}{\varepsilon + 1}\chi_t + \gamma\frac{\langle p, z\rangle}{\varepsilon + 1}\right). \tag{5.18}$$

Now, if $\langle p, z\rangle < 0$, we readily infer that $\chi_t = x_{2,t}$ converges linearly to 0 at a rate of $\chi_t \sim \exp(-|\langle p, z\rangle|/(1+\varepsilon) \cdot \gamma t)$ as predicted by Theorem 2. In particular, if $g(0) = -p = (\nu^*, \nu^*)$ 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 $g(0) = -p = (\nu^*, 0)$, 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 g is not sharp at x^*. ♦

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 5. *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*

$$\nabla h(x) = \sum_{i=1}^{n} \theta'(x_i) e_i \quad \text{for all } x \in \mathcal{X}_h. \tag{5.19}$$

Proof. See [37, Thm 23.8], whose conditions are satisfied because $\mathcal{X}$ is polyhedral. ■

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

Lemma 6. *Let $\mathcal{X}$ be a polyhedral domain of the form (5.6). 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 6 is based on Farkas' lemma so we relegate it to Appendix A. Armed with all this, we can finally proceed to prove our result for linearly constrained problems.

Proof of Theorem 2. We will consider two main cases, namely $\lim_{x\to 0^+} h'(x) = -\infty$ (the steep case) and $\lim_{x\to 0^+} h'(x) > -\infty$ (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. Note first 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:

$$\text{Under Assumption 4}(b)\text{:} \quad \theta'(x) \geq \log x - C \tag{5.20a}$$

$$\text{Under Assumption 4}(c)\text{:} \quad \theta'(x) \geq -Cx^{-p} \tag{5.20b}$$

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

- $\mathcal{B}_r \subseteq \mathcal{U} \cap \mathcal{B}$ with $\mathcal{B}$ and $\mathcal{U}$ defined by (SOS) and (4.4) respectively.
- If $x \in \mathcal{B}_r$ then $x_i < \delta$ for all $i \in \mathcal{A}$.
- If $x \in \mathcal{B}_r$, then $\|g(x) - g(x^*)\|_* \leq \nu^*/(2P)$ with ν^* by (5.11) and P by Lemma 6.

Now, recall that Step 1 of the proof of Theorem 1 implies that the iterates of (BPM) will remain in $\mathcal{B}_r$ for all t if x_1 is initialized sufficiently close to x^*. Given this stability guarantee, we will construct below two sets $\mathcal{I}_\sharp \subseteq \mathcal{A}_\sharp$, $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$ such that, for all $i \in \mathcal{I} \coloneqq \mathcal{I}_\sharp \cup \mathcal{I}_\flat$, we have

$$x_{i,t} \leq P^{|\mathcal{I}|} \cdot \begin{cases} \exp(C + P\|\nabla h(x_1)\|_* + PR - \gamma\nu_{\mathrm{eff}}(t-1)/2) & \text{under Assumption 4}(b), \\ C[\gamma\nu_{\mathrm{eff}}(t-1)/2 - P\|\nabla h(x_1)\|_* - PR]_+^{-\frac{1}{p}} & \text{under Assumption 4}(c), \end{cases} \tag{5.21}$$

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 6 and ν_{eff} is defined as in (5.14) with $\varrho = P|\mathcal{A}|$.

Our construction proceeds inductively, starting with $\mathcal{I}_\sharp = \mathcal{I}_\flat = \varnothing$, for which the stated property holds trivially. For the inductive step, if (5.21) holds for $\mathcal{I}_\sharp \subsetneq \mathcal{A}_\sharp$ and $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$, we will show that there exists some $j \in \mathcal{A} \setminus \mathcal{I}$ such that (5.21) still holds for $\mathcal{I} \cup \{j\}$. Since the number of active constraints is finite, the progressive addition of these indices will allow us to reach $\mathcal{I}_\sharp = \mathcal{A}_\sharp$, thus proviing our initial claim.

To carry all this out, assume that $\mathcal{I}_\sharp \subsetneq \mathcal{A}_\sharp$, $\mathcal{I}_\flat \subseteq \mathcal{A}_\flat$, and apply Lemma 6 to $\mathcal{I} = \mathcal{I}_\sharp \cup \mathcal{I}_\flat$. If the first case of Lemma 6 holds, then $\mathcal{I} \neq \varnothing$ and there exists $i \in \mathcal{A}$ such that

$$x_{i,t} \leq P \max\nolimits_{j\in\mathcal{I}} x_{j,t} \tag{5.22}$$

so (5.21) still holds when i is appended to $\mathcal{I}_\sharp$ or $\mathcal{I}_\flat$. Otherwise, the second case of Lemma 6 holds and there exists $z \in \ker A$ 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 $x_{t+1} = P_{x_t}(-\gamma g_{t+1/2})$ belongs to $\mathcal{X}_h = \operatorname{ri}\mathcal{X}$ for all $t = 1, 2, \dots$, so $\mathrm{NC}(x_{t+1})$ is the affine hull of $\mathcal{X}$, i.e., $\mathrm{NC}(x_{t+1}) = \operatorname{row}(A)$. Hence, Lemma A.2 guarantees

$$\nabla h(x_{t+1}) - \nabla h(x_t) + \gamma g_{t+1/2} \in \operatorname{row}(A) \tag{5.23}$$

so, telescoping from $s = 1$ to $t-1$, we get $\nabla h(x_t) - \nabla h(x_1) + \gamma \sum_{s=1}^{t-1} g_{s+1/2} \in \operatorname{row}(A)$. Thus, taking the scalar product with $z \in \ker A$ yields

$$\sum_{i=1}^{n} \theta'(x_{i,t})z_i = \langle \nabla h(x_1), z\rangle - \gamma \sum_{s=1}^{t-1} \langle g_{s+1/2}, z\rangle \tag{5.24}$$

so, after rearranging and invoking (5.8) to write $\langle g(x^*), z\rangle = \sum_{i\in\mathcal{A}} \nu_i z_i$, we get

$$\begin{aligned} \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \theta'(x_{i,t})z_i + \sum_{i\in\mathcal{I}} \theta'(x_{i,t})z_i &= \langle \nabla h(x_1), z\rangle - \gamma \sum_{s=1}^{t-1} \sum_{i\in\mathcal{A}} \nu_i z_i \\ &\quad + \gamma \sum_{s=1}^{t-1} \langle g(x^*) - g_{s+1/2}, z\rangle - \sum_{i\notin\mathcal{A}} \theta'(x_{i,t})z_i. \end{aligned} \tag{5.25}$$

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

$$\begin{aligned} \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \theta'(x_{i,t})z_i &\leq P\|\nabla h(x_1)\|_* - \gamma(t-1) \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \nu_i z_i \\ &\quad + \gamma \sum_{s=1}^{t-1} P\|g(x^*) - g_{s+1/2}\|_* + P\left\|\sum_{i\notin\mathcal{A}} \theta'(x_{i,t})e_i\right\|_* \\ &\leq P\|\nabla h(x_1)\|_* - \gamma(t-1) \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \nu_i z_i + \tfrac{1}{2}\gamma\nu^*(t-1) + PR. \end{aligned} \tag{5.26}$$

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

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

Case 1.1: 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 (5.26) gives

$$\sum_{i \in \mathcal{A} \setminus \mathcal{I}} \theta'(x_{i,t}) z_i + \gamma(t-1)\nu^* z_i \leq P\|\nabla h(x_1)\|_* + \tfrac{1}{2}\gamma\nu^*(t-1) + PR. \tag{5.27}$$

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

$$(\theta'(x_{j,t}) + \gamma(t-1)\nu^*)(|\mathcal{A}| \setminus \mathcal{I}) z_j \leq P\|\nabla h(x_1)\|_* + \tfrac{1}{2}\gamma\nu^*(t-1) + PR \tag{5.28}$$

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

$$\theta'(x_{j,t}) \leq P\|\nabla h(x_1)\|_* - \tfrac{1}{2}\gamma\nu^*(t-1) + PR\,. \tag{5.29}$$

Case 1.2: 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 (5.26) above becomes,

$$\sum_{i \in \mathcal{A} \setminus \mathcal{I}} \theta'(x_{i,t}) z_i \leq P\|\nabla h(x_1)\|_* - \tfrac{1}{2}\gamma\nu^*(t-1) + PR\,. \tag{5.30}$$

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

$$\theta'(x_{j,t}) z_j \leq P\|\nabla h(x_1)\|_* - \frac{\gamma\nu^*}{2|\mathcal{A}|}(t-1) + PR\,. \tag{5.31}$$

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

$$\theta'(x_{j,t}) \leq P\|\nabla h(x_1)\|_* - \frac{\gamma\nu^*}{2P|\mathcal{A}|}(t-1) + PR \tag{5.32}$$

Thus, in both cases, there exists $j \in \mathcal{A} \setminus \mathcal{I}$ such that (5.32) holds, since $P|\mathcal{A}| \geq 1$. Therefore, combining this inequality with (5.20) we conclude that (5.21) 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 direct. 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 $\mathcal{U} \cap \mathcal{B}$, and if $x \in \mathcal{B}_r$, then $\|g(x) - g(x^*)\|_* \leq \frac{\nu^*}{3P}$,
- 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 6. This is possible since ∇h is continuous at x^*,
- No other constraint $x_i = 0$ with $i \notin \mathcal{A}$ becomes active in $\mathcal{B}_r$.

As we have seen in the proof of Theorem 1, if (BPM) is initialized 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

$$\text{for all } i \in \mathcal{I}_\sharp \cup \mathcal{I}_\flat, x_{i,t} = 0\,. \tag{5.33}$$

Starting with $\mathcal{I}_\sharp = \mathcal{I}_\flat = \varnothing$, Eq. (5.33) 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 6 with $\mathcal{I} \coloneqq \mathcal{I}_\sharp \cup \mathcal{I}_\flat$. If the first case of Lemma 6 holds, then $\mathcal{I} \neq \varnothing$ and there exists $j \in \mathcal{A}$ such that,

$$x_{j,t} \leq P \max(x_{i,t} : i \in \mathcal{I}) \tag{5.34}$$

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 of $\mathcal{I}$ can become active, and these constraints are actually active by the definition of $\mathcal{I}$ and Eq. (5.33). Hence, the normal cone at x_t (see Eq. (5.7)) becomes

$$\mathrm{NC}(x_t) = \left\{ -\sum_{i \in \mathcal{I}} \nu_i e_i : (\nu_i)_{i \in \mathcal{I}} \in (\mathbb{R}_+)^{\mathcal{I}} \right\} + \mathrm{row}(A)\,. \tag{5.35}$$

Taking a scalar product between the last inclusion of Lemma A.2 and z, we get that,

$$\left\langle \nabla h(x_t) - \nabla h(x_{t-1}) - \gamma g_{t-1/2}, z \right\rangle = 0\,. \tag{5.36}$$

This means, from the definition of $\mathcal{B}_r^{\mathcal{X}}(x^*)$, that,

$$\gamma \langle g(x^*), z \rangle = \langle \nabla h(x_t) - \nabla h(x_{t-1}), z \rangle + \gamma \langle g(x^*) - \gamma g_{t-1/2}, z \rangle \leq \frac{2\gamma\nu^*}{3} \tag{5.37}$$

However, by (5.8) and the properties of z, we also have

$$\langle g(x^*), z \rangle = \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \nu_i z_i \geq \nu^* \tag{5.38}$$

which is in contradiction with (5.37). We may therefore iteratively add coordinates of $\mathcal{A}$ for which $x_{i,t} = 0$; this 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 [28] – which concerns the method's ergodic average and advocates the use of non-Euclidean regularizers in domains with a favorable geometry – it *does* run contrary to its spirit. We attribute the source of this discrepancy 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 [2, 6, 22]. 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). This work has also been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Part of this work was done while P. Mertikopoulos was visiting the Simons Institute for the Theory of Computing.

Appendix A. Auxiliary results

We provide here a series of basic properties, helper lemmas and auxiliary results that we use repeatedly in our paper.

A.1. **Lemmas on numerical sequences.**

The first two results concern numerical sequences.

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

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

Then, for all $t = 1, 2, \ldots,$ 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 [34, p. 46]. ■

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

Lemma 1. *Suppose that $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ admits the asymptotic expansion*

$$f(x) = x - \lambda x^{1+r} + o(x^{1+r}) \quad \text{as } x \to 0 \tag{3.7}$$

for positive constants $\lambda, r > 0$. Then, for $u_1 > 0$ small enough, the sequence $u_{t+1} = f(u_t)$, $t = 1, 2, \ldots,$ 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

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

Note first that, if $u_1 \leq \varepsilon$, Lemma A.1 readily implies that u_t converges to 0 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, we consider the sequence u_t^{-r}, $t = 1, 2, \ldots,$ for which we get

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

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

A.2. **Properties of Bregman divergences and the induced prox-mappings.**

We recall some basic properties of the Bregman divergence and the induced prox-mapping. Variants of these properties are fairly well known in the literature, so we omit their proofs and we refer to [7, 18, 25].

Lemma A.2. *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) $\partial h(x) = \nabla h(x) + \mathrm{PC}(x)$ (A.5a)

b) $x^+ = P_x(y) \iff \nabla h(x) + y \in \partial h(x^+) \iff \nabla h(x^+) - \nabla h(x) \in y - \mathrm{PC}(x^+)$ (A.5b)

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.

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

$$D(p, x^+) = D(p, x) + D(x, x^+) + \langle \nabla h(x^+) - \nabla h(x), x - p\rangle \tag{A.6}$$

Lemma A.4 (Non-expansiveness)**.** *For all $x \in \mathcal{X}_h$ and all $y, y^+ \in \mathcal{Y}$ we have:*

$$\|P_x(y^+) - P_x(y)\| \leq \|y^+ - y\|_* \tag{A.7}$$

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

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

$$\begin{aligned} D(p, x^+) &\leq D(p, x) + \langle y - v, x^+ - p\rangle - D(x^+, x) && \text{(4.8a)} \\ &\leq D(p, x) + \langle y - v, x - p\rangle + \tfrac{1}{2}\|y - v\|_*^2 \,. && \text{(4.8b)} \end{aligned}$$

Proof. Our proof follows [26, Proposition B.3], but with a slight modification to account for the extra term with $v \in \mathrm{PC}(p)$. The first step is to invoke the three-point identity (A.6) to write

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

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

$$\begin{aligned} D(p, x^+) &= D(p, x) - D(x^+, x) - \langle \nabla h(x) - \nabla h(x^+), x^+ - p\rangle \\ &\leq D(p, x) - D(x^+, x) - \langle y, x^+ - p\rangle && \text{(A.9)} \end{aligned}$$

where the inequality in the last line follows from Lemma A.2. Hence, given that $\langle v, x^+ - p\rangle \leq 0$ by the fact that $v \in \mathrm{PC}(p)$, we readily obtain

$$D(p, x^+) \leq D(p, x) - D(x^+, x) - \langle y - v, x^+ - p\rangle \,. \tag{A.10}$$

For the second inequality of the lemma, note that

$$\begin{aligned} -\langle y - v, x^+ - p\rangle &= -\langle y - v, x^+ - x\rangle - \langle y - v, x - x^+\rangle \\ &\leq \tfrac{1}{2}\|y - v\|_*^2 + \tfrac{1}{2}\|x^+ - x\|^2 - \langle y - v, x - x^+\rangle \\ &\leq \tfrac{1}{2}\|y - v\|_*^2 + D(x^+, x) - \langle y - v, x - x^+\rangle && \text{(A.11)} \end{aligned}$$

where the penultimate inequality follows directly from Young's inequality and the last one from (4.1). Our assertion is then obtained by combining this last bound with (A.10). ■

Lemma 3. *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:*

$$D(p, x_2^+) \leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 - \tfrac{1}{2}\|x_1^+ - x\|^2. \tag{4.9}$$

Proof. Our proof follows [26, Proposotion B.4], again with a slight modification to account for the extra terms with $v \in \mathrm{PC}(p)$. Specifically, applying Lemma 2 with $x_2^+ = P_x(y_2)$ and $v \in \mathrm{PC}(p)$ gives

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) + \langle y_2 - v, x_2^+ - p\rangle - D(x_2^+, x) \\ &\leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \langle y_2 - v, x_2^+ - x_1^+\rangle - D(x_2^+, x) && \text{(A.12)} \end{aligned}$$

To lower bound $D(x_2^+, x)$, we use again Lemma 2 with $p \leftarrow x_2^+$ and $x_1^+ = P_x(y_1)$, which gives

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

and, after rearranging to isolate $D(x_2^+, x)$ and substituting the resulting bound in (A.12),

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

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

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 \\ &\quad + \tfrac{1}{2}\|x_2^+ - x_1^+\|^2 - \tfrac{1}{2}\|x_2^+ - x_1^+\|^2 - \tfrac{1}{2}\|x_1^+ - x\|^2 \\ &\leq D(p, x) + \langle y_2 - v, x_1^+ - p\rangle + \tfrac{1}{2}\|y_2 - y_1 - v\|_*^2 - \tfrac{1}{2}\|x_1^+ - x\|^2 \end{aligned} \tag{A.15}$$

and our proof is complete. ■

A.3. **Legendre exponent for interior points.** We discussed in Section 4.1 that $\beta_h(p) = 0$ whenever p is an interior point. We formalize this intuition below.

Lemma A.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

$$\|\nabla h(p) - \nabla h(x)\|_* \leq \kappa\|p - x\| \quad \text{for all } x \in \mathcal{U} \cap \mathcal{X}_h. \tag{A.16}$$

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

$$\begin{aligned} D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x\rangle &\leq \langle \nabla h(p) - \nabla h(x), p - x\rangle \\ &\leq \|\nabla h(p) - \nabla h(x)\|_*\|p - x\| \leq \kappa\|p - x\|^2 \end{aligned} \tag{A.17}$$

for all $x \in \mathcal{U} \cap \mathcal{X}_h$. This shows that (4.3) holds with $\beta = 0$, i.e., $\beta_h(p) = 0$. ■

A.4. **A separation result.** We now prove Lemma 6, which we restate below for convenience:

Lemma 6. *Let $\mathcal{X}$ be a polyhedral domain of the form (5.6). 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

$$\mathcal{X} \subset \left\{ x \in \mathbb{R}^n : \sum_{i \in \mathcal{A} \setminus \mathcal{I}} \lambda_i x_i = \sum_{i \in \mathcal{I}} \mu_i x_i \right\} \tag{A.18}$$

holds. Case (a) considers the case when such coefficients exist, while Case (b) 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.18) 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 (5.6) of $\mathcal{X}$. In addition, there is some $i \in \mathcal{A}\setminus\mathcal{I}$ such that $\lambda_i > 0$ and thus we have

$$\forall x \in \mathcal{X},\, x_i \leq \frac{\max\left(|\lambda_j| : j \in \mathcal{I}\right)}{\lambda_i} \max(x_j : j \in \mathcal{I}) \tag{A.19}$$

which corresponds to the first case of the lemma.

Case (b). 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.18) does not hold. To interpret this situation, we use the fact that $\mathcal{X}$ is of the general polyhedral form (5.6) 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

$$\begin{aligned}
\text{Eq. (A.18)} \iff\ & \operatorname{aff}\mathcal{X} \subset \left\{ x \in \mathbb{R}^n : \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \lambda_i x_i = \sum_{i\in\mathcal{I}} \mu_i x_i \right\} \\
\iff\ & \ker A \subset \left\{ x \in \mathbb{R}^n : \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \lambda_i x_i = \sum_{i\in\mathcal{I}} \mu_i x_i \right\} \\
\iff\ & \sum_{i\in\mathcal{A}\setminus\mathcal{I}} \lambda_i e_i - \sum_{i\in\mathcal{I}} \mu_i e_i \in \operatorname{row}(A).
\end{aligned} \tag{A.20}$$

Therefore, the fact that (A.18) 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,

$$\sum_{i\in\mathcal{A}\setminus\mathcal{I}} \lambda_i e_i = \sum_{i\in\mathcal{I}} \mu_i e_i + A^\top r\,, \tag{A.21}$$

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])[4], this means that the system,

$$\begin{cases} z_i > 0 & \text{for } i \in \mathcal{A}\setminus\mathcal{I} \\ z_i = 0 & \text{for } i \in \mathcal{A} \\ Az = 0 \end{cases} \tag{A.22}$$

admits a solution $z \in \mathbb{R}^n$. Rescaling 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. 6)**.** We proceed to compute the Taylor expansion of F near $x^* = -1$ for the shifted operator $g(x) = x + 1$. Indeed, in this case, the fixed point operator F is given by

$$F(x) = P_x(-\gamma g(x)) = P_x(-\gamma(x+1))$$

[4]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.

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

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

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

Another Taylor expansion then yields

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

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

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

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

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

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

Example 5.4 (Three-dimensional simplex, continuing from p. 15)**.** 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 $g_2(x_t) = x_{2,t} = o(1)$ so, arguing as in the first part of the example, we readily get

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

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

$$\begin{aligned}\frac{x_{3,t+1}}{x_{2,t+1}} &= \frac{x_{3,t}}{x_{2,t}}\exp(\gamma g_2(x_t) - \gamma g_3(x_t)) = \frac{x_{3,t}}{x_{2,t}}\exp(\gamma x_{2,t} - \gamma(x_{3,t}-1)) \\ &= \frac{x_{3,t}}{x_{2,t}}\exp(2\gamma x_{2,t} + \gamma x_{1,t}) \leq \frac{x_{3,t}}{x_{2,t}}\exp\left(2\gamma\frac{x_{2,t}}{x_{3,t}} + \gamma x_{1,t}\right) \end{aligned} \tag{B.7}$$

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

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

since $x_{1,t}/x_{2,t}$ vanishes as $t \to \infty$. Hence, after telescoping, we conclude that $\frac{x_{3,t}}{x_{2,t}} \leq 2\gamma t + o(t)$, which in turn shows that $x_{2,t} \sim x_{2,t}/x_{3,t} = \Omega(1/t)$, as claimed. ♦

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] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[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] Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[10] Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015.

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

[12] 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 Conference on Learning Theory*, 2012.

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

[14] Fudenberg, D. and Tirole, J. *Game Theory.* The MIT Press, 1991.

[15] 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.

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

[17] 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.

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

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

[20] Kiwiel, K. C. Proximal minimization methods with generalized Bregman functions. *SIAM Journal on Control and Optimization*, 35:1142–1168, 1997.

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

[22] 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.

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

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

[25] 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.

[26] 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.

[27] Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. In *AISTATS '20: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[28] 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.

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

[30] 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.

[31] Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course.* Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

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

[33] Pauwels, E. On the nature of bregman functions. *arXiv preprint arXiv:2302.02689*, 2023.

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

[35] 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.

[36] 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.

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

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

[39] Solan, E. and Vieille, N. Stochastic games. *Proceedings of the National Academy of Sciences*, 112 (45):13743–13746, 2015.

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

[41] Tsuda, K., Rätsch, G., and Warmuth, M. K. Matrix exponentiated gradient updates for on-line Bregman projection. *Journal of Machine Learning Research*, 6:995–1018, 2005.

[42] Vanderbei, R. J., Meketon, M. S., and Freedman, B. A. A modification of Karmarkar's linear programming algorithm. *Algorithmica*, 1(1):395–407, November 1986.

[43] Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.