Proceedings of Machine Learning Research vol 134:1–31, 2021

The Last-Iterate Convergence Rate of Optimistic Mirror Descent in Stochastic Variational Inequalities

Waïss Azizian WAISS.AZIZIAN@ENS.FR

Franck Iutzeler FRANCK.IUTZELER@UNIV-GRENOBLE-ALPES.FR

Jérôme Malick JEROME.MALICK@UNIV-GRENOBLE-ALPES.FR

Univ. Grenoble Alpes, LJK, 38000, Grenoble, France

Panayotis Mertikopoulos PANAYOTIS.MERTIKOPOULOS@IMAG.FR

Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000, Grenoble, France & Criteo AI Lab

Abstract

In this paper, we analyze the convergence rate of optimistic mirror descent methods in stochastic variational inequalities, a class of optimization problems with important applications to learning theory and machine learning. Our analysis reveals an intricate relation between the algorithm's rate of convergence and the local geometry induced by the method's underlying Bregman function. We quantify this relation by means of the *Legendre exponent*, a notion that we introduce to measure the growth rate of the Bregman divergence relative to the ambient norm near a solution. We show that this exponent determines both the optimal step-size policy of the algorithm and the optimal rates attained, explaining in this way the differences observed for some popular Bregman functions (Euclidean projection, negative entropy, fractional power, etc.).

1. Introduction

Variational inequalities – and, in particular, *stochastic* variational inequalities – have recently attracted considerable attention in machine learning and learning theory as a flexible paradigm for "optimization beyond minimization" – i.e., for problems where finding an optimal solution does not necessarily involve minimizing a loss function. In this context, our paper examines the rate of convergence of *optimistic mirror descent* (OMD), a state-of-the-art algorithmic template for solving variational inequalities (VIs) that incorporates an "optimistic" look-ahead step with a "mirror descent" apparatus relying on a suitably chosen Bregman kernel. Our contributions focus exclusively on the stochastic case, which is of central interest to learning theory. To put them in context, we begin with a general overview below and discuss more specialized references in Section 5.

General overview. Algorithms for solving variational inequalities have a very rich history in optimization; for a survey, see Facchinei and Pang (2003). At a high level, if the vector field defining the problem is bounded and strictly monotone, simple forward-backward schemes are known to converge – and if combined with a Polyak–Ruppert averaging scheme, they achieve an $\mathcal{O}(1/\sqrt{t})$ rate of convergence without the strictness caveat (Bruck Jr., 1977; Passty, 1979). If, in addition, the problem's defining vector field is Lipschitz continuous, the extra-gradient (EG) algorithm of Korpelevich (1976) achieves convergence without strict monotonicity requirements, while the method's ergodic average converges at a $\mathcal{O}(1/t)$ rate (Nemirovski, 2004).

In the stochastic case, the landscape is considerably different. For VI problems with a monotone operator, the stochastic version of the EG algorithm achieves an ergodic $\mathcal{O}(1/\sqrt{t})$ convergence rate (Juditsky et al., 2011). If the problem is *strongly* monotone, this rate improves to $\mathcal{O}(1/t)$ and it is achieved by simple forward-backward methods, with or without averaging (Hsieh et al., 2019; Nemirovski et al., 2009).

In this context, the optimistic mirror descent algorithm has been designed to meet two complementing objectives: (*i*) improve the dependence of the above rates on the problem's dimensionality; and (*ii*) minimize the number of oracle queries per iteration. The first of these objectives is achieved by the "mirror descent" component: by employing a suitable *distance-generating function* (DGF) – like the negative entropy on the simplex – mirror descent achieves convergence rates that are (almost) dimension-free in problems with a favorable geometry. This idea dates back to Nemirovski and Yudin (1983), and it is also the main building block of the mirror-prox (MP) algorithm which achieves order-optimal rates with two oracle queries per iteration (Nemirovski, 2004).

The "optimistic" module then clicks on top of it by replacing one of the two queries by an already observed gradient. This "information reuse" idea was originally due to Popov (1980), and it has recently resurfaced several times in learning theory, cf. Chiang et al. (2012), Rakhlin and Sridharan (2013a,b), Daskalakis et al. (2018), Gidel et al. (2019) and Hsieh et al. (2019).

Our contributions. The aim of our paper is to examine the convergence rate of OMD in stochastic VI problems. For generality, we focus on non-monotone VIs, and we investigate the convergence to local solutions that satisfy a second-order sufficient condition.

In this regard, our first finding is that the algorithm's rate of convergence depends sharply on the local geometry induced by the underlying Bregman function. We formalize this by introducing the notion of the *Legendre exponent*, which can roughly be described as the logarithmic ratio of the volume of a regular Euclidean ball centered at the solution under study to that of a Bregman ball of the same radius. For example, the ordinary Euclidean version of OMD has a Legendre exponent of $\beta = 0$; on the other hand, the entropic variant – which has important applications to learning and game theory due to its connection to the exponential weights algorithm, cf. Daskalakis and Panageas (2019) – has a Legendre exponent of $\beta = 1/2$ on boundary points.

As a function of β, we obtain the following rates:

$$\mathcal{O}\Big(t^{-(1-\beta)}\Big) \text{ if } 0 \leq \beta < 1/2 \qquad \text{and} \qquad \mathcal{O}\Big(t^{-\frac{1-\varepsilon}{2\beta}(1-\beta)}\Big) \text{ if } 1/2 \leq \beta < 1$$

for arbitrary small $\varepsilon > 0$. Interestingly, these guarantees undergo a first-order phase transition between the Euclidean-like phase ($0 \leq \beta < 1/2$) and the entropy-like phase ($1/2 \leq \beta < 1$). This coincides with the dichotomy between steep and non-steep DGFs (i.e., DGFs that are differentiable only on the interior of the problem's domain versus those that are differentiable throughout). Moreover, these rate guarantees are reached for different step-size schedules, respectively $\gamma_t = \Theta(1/t^{1-\beta})$ and $\Theta(1/t^{(1-\varepsilon)/2})$, also depending on the Legendre exponent.

To the best of our knowledge, the only comparable result in the literature is the recent work of Hsieh et al. (2019) that derives an $\mathcal{O}(1/t)$ convergence rate for the Euclidean case (i.e., when $\beta = 0$). Along with some other recent works on optimistic gradient methods, we discuss this in detail in Section 5, after the precise statement of our results.

2. Problem setup and preliminaries

Problem formulation, examples, and blanket assumptions. Let $\mathcal{X}$ be a d-dimensional real space with norm $\|\cdot\|$, and let $\mathcal{K}$ be a closed convex subset thereof. Throughout what follows, we will focus on solving (Stampacchia) variational inequalities of the general form:

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

In the above, $\langle y, x \rangle$ denotes the canonical pairing between $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ and the *defining vector field* $v \colon \mathcal{K} \to \mathcal{Y}$ of the problem is a single-valued operator with values in the dual space $\mathcal{Y} \coloneqq \mathcal{X}^*$ of $\mathcal{X}$. Letting $\|y\|_* \coloneqq \max\{\langle y, x \rangle : \|x\| \leq 1\}$ denote the induced dual norm on $\mathcal{Y}$, we will make the following blanket assumption for v.

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

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

For concreteness, we provide below some archetypal examples of VI problems.

Example 2.1. Consider the minimization problem

$$\text{minimize}_{x \in \mathcal{K}} \quad f(x) \tag{min}$$

with $f \colon \mathcal{K} \to \mathbb{R}$ assumed C^1-smooth. Then, letting $v(x) = \nabla f(x)$, the solutions of (VI) are precisely the Karush–Kuhn–Tucker (KKT) points of (min), cf. Facchinei and Pang (2003). ◀

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

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

where $\mathcal{K}_1 \subseteq \mathbb{R}^{d_1}$ and $\mathcal{K}_2 \subseteq \mathbb{R}^{d_2}$ are convex and closed, and the problem's objective function $\Phi \colon \mathcal{K}_1 \times \mathcal{K}_2 \to \mathbb{R}$ is again assumed to be smooth. In the game-theoretic interpretation of von Neumann (1928), x_1 is controlled by a player seeking to minimize $\Phi(\cdot, x_2)$, whereas x_2 is controlled by a player seeking to maximize $\Phi(x_1, \cdot)$. Accordingly, solving (SP) consists of finding a *min-max point* $(x_1^*, x_2^*) \in \mathcal{K} \coloneqq \mathcal{K}_1 \times \mathcal{K}_2$ such that

$$\Phi(x_1^*, x_2) \leq \Phi(x_1^*, x_2^*) \leq \Phi(x_1, x_2^*) \quad \text{for all } x_1 \in \mathcal{K}_1, x_2 \in \mathcal{K}_2. \tag{min-max}$$

Min-max points may not exist if Φ is not convex-concave. In this case, one looks instead for first-order stationary points of Φ, i.e., action profiles $(x_1^*, x_2^*) \in \mathcal{K}_1 \times \mathcal{K}_2$ such that x_1^* is a KKT point of $\Phi(\cdot, x_2^*)$ and x_2^* is a KKT point of $-\Phi(x_1^*, \cdot)$. Letting $x = (x_1, x_2)$ and $v = (\nabla_{x_1}\Phi, -\nabla_{x_2}\Phi)$, we see that the solutions of (VI) are precisely the first-order stationary points of Φ. ◀

The above examples show that not all solutions of (VI) are desirable: for example, such a solution could be a local *maximum* of f in the case of (min) or a max-min point in the case of (min-max). For this reason, we will concentrate on solutions of (VI) that satisfy the following second-order condition:

Assumption 2 (Second-order sufficiency)**.** For a solution x^* of (VI), there exists a neighborhood $\mathcal{B}$ and a positive constant $\mu > 0$ such that

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

In the context of (min), Assumption 2 implies that f grows (at least) quadratically along every ray emanating from x^*; in particular, we have $f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle + (\mu/2)\|x - x^*\|^2 = \Omega(\|x - x^*\|^2)$ for all $x \in \mathcal{B}$ (though this does not mean that f is strongly convex in $\mathcal{B}$). Likewise, in the context of (min-max), Assumption 2 gives $\Phi(x_1, x_2^*) = \Omega(\|x_1 - x_1^*\|^2)$ and $\Phi(x_1^*, x_2) = -\Omega(\|x_2 - x_2^*\|^2)$, so x^* is a local Nash equilibrium of Φ. In general, Assumption 2 guarantees that x^* is the unique solution of (VI) in $\mathcal{B}$. (Indeed, any other solution $\hat{x} \neq x^*$ of (VI) would satisfy $0 \geq \langle v(\hat{x}), \hat{x} - x^* \rangle \geq \mu \|\hat{x} - x^*\|^2 > 0$, a contradiction.)

3. The optimistic mirror descent algorithm

In this section, we recall the *optimistic mirror descent* (OMD) method for solving (VI). To streamline the flow of our paper, we first discuss the type of feedback available to the optimizer.

3.1. The oracle model

OMD is a first-order method that only requires access to the problem's defining vector field via a "black-box oracle" that returns a (possibly imperfect) version of $v(x)$ at the selected query point $x \in \mathcal{K}$. Concretely, we focus on *stochastic first-order oracles* of the form

$$\mathsf{V}(x; \theta) = v(x) + \mathsf{Err}(x; \theta) \tag{SFO}$$

where θ is a random variable taking values in some abstract measure space Θ and $\mathsf{Err}(x; \theta)$ is an umbrella error term capturing all sources of uncertainty in the model.

In practice, the oracle is called repeatedly at a sequence of query points $X_t \in \mathcal{K}$ with a different random seed θ_t at each time.[1] The sequence of oracle signals $V_t = \mathsf{V}(X_t; \theta_t)$ may then be written as

$$V_t = v(X_t) + U_t \tag{1}$$

where $U_t = \mathsf{Err}(X_t; \theta_t)$ denotes the error of the oracle model at stage t. To keep track of this sequence of events, we will treat X_t as a stochastic process on some complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and we will write $\mathcal{F}_t \coloneqq \mathcal{F}(X_1, \dots, X_t) \subseteq \mathcal{F}$ for the *history of play* up to stage t (inclusive). Since the randomness entering the oracle is triggered only *after* a query point has been selected, we will posit throughout that θ_t (and hence U_t and V_t) is $\mathcal{F}_{t+1}$-measurable – but not necessarily $\mathcal{F}_t$-measurable. We will also make the blanket assumption below for the oracle.

Assumption 3 (Oracle signal)**.** The error term U_t in (1) satisfies the following properties:

$$a) \quad \textit{Zero-mean:} \qquad \mathbb{E}[U_t \,|\, \mathcal{F}_t] = 0. \tag{2a}$$

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

Both assumptions are standard in the literature on stochastic methods in optimization, cf. Polyak (1987), Hazan (2012), Bubeck (2015), and references therein.

1. In the sequel, we will allow the index t to take both integer and half-integer values.

3.2. Optimistic mirror descent

With all this in hand, the optimistic mirror descent algorithm is defined in recursive form as follows:

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

In the above, $t = 1, 2, \dots$, is the algorithm's iteration counter, X_t and $X_{t+1/2}$ respectively denote the algorithm's *base* and *leading* states at stage t, $V_{t+1/2}$ is an oracle signal obtained by querying (SFO) at $X_{t+1/2}$, and $P_x(y)$ denotes the method's so-called "prox-mapping". In terms of initialization, we also take $X_1 = X_{1/2} \in \operatorname{ri} \mathcal{K}$ for simplicity. All these elements are defined in detail below.

At a high level, (OMD) seeks to leverage past feedback to anticipate the landscape of v and perform more informed steps in subsequent iterations. In more detail, starting at some base state X_t, $t = 1, 2, \dots$, the algorithm first generates an intermediate, leading state $X_{t+1/2}$, and then updates the base state with oracle input from $X_{t+1/2}$ – that is, $V_{t+1/2}$. This is also the main idea behind the extra-gradient algorithm of Korpelevich (1976); the key difference here is that OMD avoids making two oracle queries per iteration by using the oracle input $V_{t-1/2}$ received at the previous leading state $X_{t-1/2}$ as a proxy for V_t (which is never requested or received by the optimizer).

The second basic component of the method is the prox-mapping P, which, loosely speaking, can be viewed as a generalized, proximal/projection operator adapted to the geometry of the problem's primitives. To define it formally, we will need the notion of a *distance-generating function* on $\mathcal{K}$:

Definition 1. A convex lower semi-continuous function $h\colon \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ (with $\operatorname{dom} h = \mathcal{K}$) is a *distance-generating function* (DGF) on $\mathcal{K}$ if

1. 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 $x \in \operatorname{dom} \partial h$.
2. h is continuous and 1-strongly convex on $\mathcal{K}$, i.e., for all $x \in \operatorname{dom} \partial h, x' \in \operatorname{dom} h$, we have:

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2. \tag{3}$$

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

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x \rangle, \qquad \text{for all } p \in \mathcal{K}, x \in \mathcal{K}_h \tag{4}$$

and the induced *prox-mapping* $P\colon \mathcal{K}_h \times \mathcal{Y} \to \mathcal{K}_h$ as

$$P_x(y) = \operatorname*{arg\,min}_{x' \in \mathcal{K}} \{\langle y, x - x' \rangle + D(x', x)\} \qquad \text{for all } x \in \mathcal{K}_h, y \in \mathcal{Y}. \tag{5}$$

In the rest of this section we take a closer look at some commonly used DGFs and the corresponding prox-mappings; for simplicity, we focus on one-dimensional problems on $\mathcal{K} = [0, 1]$.

Example 3.1 (Euclidean projection)**.** For the quadratic DGF $h(x) = x^2/2$ for $x \in \mathcal{K}$, the Bregman divergence is $D(p, x) = (p - x)^2/2$ and the induced prox-mapping is the Euclidean projector $P_x(y) = [x + y]_0^1$. The prox-domain of h is the entire feasible region, i.e., $\mathcal{K}_h = \mathcal{K} = [0, 1]$. ◀

Example 3.2 (Negative entropy)**.** Another popular example is the entropic DGF $h(x) = x\log x + (1-x)\log(1-x)$. The corresponding Bregman divergence is the relative entropy $D(p,x) = p\log\frac{p}{x} + (1-p)\log\frac{1-p}{1-x}$. A standard calculation shows that the prox-mapping is $P_x(y) = xe^y/[(1-x)+xe^y]$. In contrast to Example 3.1, the prox-domain of h is $\mathcal{K}_h = (0,1)$. ◀

Example 3.3 (Tsallis entropy)**.** An alternative to the Gibbs–Shannon negentropy is the *Tsallis / fractional power* DGF $h(x) = -\frac{1}{q(1-q)}\cdot[x^q + (1-x)^q]$ for $q > 0$ (Tsallis, 1988). Formally, to define h for $q \leftarrow 1$, we will employ the continuity convention $t^q/(q-1) \leftarrow \log t$, which yields the entropic DGF of Example 3.2. Instead, for $q \leftarrow 2$, we readily get the Euclidean DGF of Example 3.1. The corresponding prox-mapping does not admit a closed-form expression for all p, but it can always be calculated in logarithmic time with a simple binary search. Finally, we note that the prox-domain $\mathcal{K}_h$ of h is the entire space $\mathcal{K}$ for $q > 1$; on the other hand, for $q \in (0,1]$, we have $\mathcal{K}_h = (0,1)$. ◀

The above examples will play a key role in illustrating the analysis to come and we will use them as running examples throughout.

4. The geometry of distance-generating functions

The Bregman topology. To proceed with our analysis and the statement of our results for (OMD), we will need to take a closer look at the geometry induced on $\mathcal{K}$ by the choice of h. The first observation is that, by the strong convexity requirement for h, we have:

$$D(p,x) \geq \tfrac{1}{2}\|p-x\|^2 \quad \text{for all } p \in \mathcal{K},\ x \in \mathcal{K}_h. \tag{6}$$

Topologically, this means that the convergence topology induced by the Bregman divergence of h on $\mathcal{K}$ is *at least as fine* as the corresponding norm topology: if a sequence x_t converges to $p \in \mathcal{K}$ in the Bregman sense ($D(p,x_t) \to 0$), it also converges in the ordinary norm topology ($\|x_t - p\| \to 0$).

On the other hand, the converse of this statement is, in general, false: specifically, the norm topology could be *strictly coarser* than the Bregman topology. To see this, let $\mathcal{K}$ be the unit Euclidean ball in $\mathbb{R}^d$, i.e., $\mathcal{K} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. For this space, a popular choice of DGF is the *Hellinger regularizer* $h(x) = -\sqrt{1-\|x\|_2^2}$, which has $\mathcal{K}_h = \operatorname{ri}\mathcal{K} = \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$ Alvarez et al. (2004); Bauschke et al. (2017). For all $\|p\|_2 = 1$, the Bregman divergence is

$$D(p,x) = \frac{1 - \sum_{i=1}^d p_i x_i}{\sqrt{1-\|x\|_2^2}}. \tag{7}$$

Shadowing the Euclidean definition, a "Hellinger sphere" of radius r centered at p is defined as the corresponding level set of $D(p,x)$, i.e., $S_p(r) = \{x \in \mathcal{K}_h : D(p,x) = r\}$. Hence, by (7), this means that a point $x \in \mathcal{K}_h$ is at "Hellinger distance" r relative to p if

$$1 - \sum\nolimits_i p_i x_i = r\sqrt{1-\|x\|_2^2}. \tag{8}$$

Now, if we take a sequence x_t converging to p in the ordinary Euclidean sense (i.e., $\|x_t - p\|_2 \to 0$), both sides of (8) will converge to 0. This has several counterintuitive consequences: (*i*) $S_r(p)$ is *not closed* in the Euclidean topology; and (*ii*) the "Hellinger center" p of $S_r(p)$ actually belongs to the Euclidean closure of $S_r(p)$. Sequentially, this shows that we can have a sequence x_t with $\|x_t - p\|_2 \to 0$, but which remains at *constant* Hellinger divergence relative to p.

From a qualitative standpoint, this discrepancy means that $D(p, x)$ is not a reliable measure of convergence of x to p: if $D(p, x)$ is large, this does not necessarily mean that p and x are topologically "far". For this reason, in the analysis of non-Euclidean instances of Bregman proximal methods, it is common (see e.g. **?**) to make the so-called "reciprocity" assumption

$$D(p, x_t) \to 0 \quad \text{whenever} \quad x_t \to p. \tag{9}$$

A straightforward verification shows that the DGFs of Examples 3.1–3.3 all satisfy (9); by contrast, this requirement of course fails for the Hellinger regularizer above.

The Legendre exponent. From a *quantitative* standpoint, even if (9) is satisfied, the local geometry induced by a Bregman DGF could still be significantly different from the base, Euclidean case. A clear example of this is provided by contrasting the Euclidean and entropic DGFs of Examples 3.1 and 3.2. Concretely, if we focus on the base point $p = 0$, we have:

1. In the Euclidean case (Example 3.1): $D(0, x) = x^2/2 = \Theta(x^2)$ for small $x \geq 0$;
2. In the entropic case (Example 3.2): $D(0, x) = -\log(1 - x) = \Theta(x)$ for small $x \geq 0$.

Consequently, the rate of convergence of x_t to $p = 0$ is drastically different if it is measured relative to the Euclidean divergence of Example 3.1 or the Kullback-Leibler divergence of Example 3.2. As an example, if we take the sequence $x_t = 1/t$, we get $D(0, x_t) = \Theta(1/t^2)$ in terms of the Euclidean divergence, but $D(0, x_t) = \Theta(1/t)$ in terms of the Kullback-Leibler divergence.

This disparity is due to the fact that, a priori, the inequality (6) cannot be inverted (even to leading order). To account for this, we introduce below the notion of the *Legendre exponent* of h.

Definition 2. Let h be a DGF on $\mathcal{K}$. Then the *Legendre exponent* of h at $p \in \mathcal{K}$ is defined to be the smallest $\beta \in [0, 1)$ such that there exists a neighborhood $\mathcal{V}$ of p and some $\kappa \geq 0$ such that

$$D(p, x) \leq \tfrac{1}{2}\kappa \|p - x\|^{2(1-\beta)} \quad \text{for all } x \in \mathcal{V} \cap \mathcal{K}_h. \tag{10}$$

Heuristically, the Legendre exponent measures the difference in relative size between ordinary "norm neighborhoods" in $\mathcal{K}$ and the corresponding "Bregman neighborhoods" induced by the sublevel sets of the Bregman divergence. It is straightforward to see that the requirement $\beta \geq 0$ is imposed by the strong convexity of h: since $D(p, x) = \Omega(\|p - x\|^2)$ for x close to p, we cannot also have $D(p, x) = \mathcal{O}(\|p - x\|^{2+\varepsilon})$ for any $\varepsilon > 0$. Note also that the notion of a Legendre exponent is only relevant if h satisfies the reciprocity requirement (9): otherwise, we could have $\sup_{x \in \mathcal{U}} D(p, x) = \infty$ for any neighborhood $\mathcal{U}$ of p, in which case we say that the Legendre exponent is 1 by convention.

To see the Legendre exponents in action, we compute them for our running examples:

1. *Euclidean projection* (Example 3.1): $D(p, x) = (p - x)^2/2$ for all $p, x \in \mathcal{K}$, so $\beta = 0$ for all p.
2. *Negative entropy* (Example 3.2): If $p \in \mathcal{K}_h = (0, 1)$, a Taylor expansion with Lagrange remainder gives $D(p, x) = h(p) - h(x) - h'(x)(p - x) = \mathcal{O}((p - x)^2)$ and thus $\beta = 0$. Now if $p = 0$, $D(0, x) = -\log(1 - x) = \Theta(x)$ which implies $\beta = 1/2$. Symmetrically, if $p = 1$, $\beta = 1/2$ too.
3. *Tsallis entropy* (Example 3.3): Assume that $q \neq 1$ (the case $q = 1$ has been treated above). We then have two cases: if $p \in \mathcal{K}_h = (0, 1)$, h is twice continuously differentiable in a neighborhood of p so that, as previously, by the Taylor formula, $D(p, x)$ On the other hand, if $p = 0$,

$$D(0, x) = \frac{x^q}{q} - \frac{1}{q(1 - q)}[1 - (1 - x)^q - q(1 - x)^{q-1}x],$$

so $D(0,x) = \Theta(x^{\min(q,2)})$ when x goes to zero. By symmetry, the situation is the same at $p = 1$, so the Legendre exponent of the Tsallis entropy is $\beta = \max(0, 1 - q/2)$.

4. *Hellinger regularizer:* As we saw at the beginning of section, the Hellinger divergence (7) does not satisfy (9), so $\beta = 1$ by convention.

Of course, beyond our running examples, the same rationale applies when $\mathcal{K}$ is a subset of a d-dimensional euclidean space. If $p \in \mathcal{K}_h$, then the associated Legendre exponent is typically 0 (see Lemma A.7 for a precise statement), while if p lies at the boundary of $\mathcal{K}$, the exponent may be positive; we illustrate this below for the important case of the simplex with the KL divergence.

Example 4.1 (Entropy on the simplex)**.** Consider the d-dimensional simplex $\mathcal{K} = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^d x_i = 1\}$ and the entropy regularizer $h(x) = \sum_{i=1}^d x_i \log x_i$. The associated divergence (also called Kullback-Leibler divergence) which can decomposed as

$$D(p,x) = \sum_{i:p_i=0} -x_i \log x_i + (\log x_i + 1)x_i + \sum_{i:p_i\neq 0} p_i \log p_i - x_i \log x_i - (\log x_i + 1)(p_i - x_i).$$

The first sum is simply $\sum_{i:p_i=0} x_i$ and the terms of the second are bounded by $\mathcal{O}((p_i - x_i)^2)$ thus

$$D(p,x) = \sum_{i:p_i=0} x_i + \mathcal{O}\left(\sum_{i:p_i\neq 0} (p_i - x_i)^2\right).$$

If $p \in \mathcal{K}_h = \{x \in \mathbb{R}^d_{++} : \sum_{i=1}^d x_i = 1\}$, we get $\beta = 0$. If $p \notin \mathcal{K}_h$ there is at least one $p_i = 0$, $D(p,x)$ is not bounded by $\mathcal{O}(\|p - x\|^2)$ but by $\mathcal{O}(\|p - x\|)$ so that $\beta = 1/2$ for such p. ◀

5. Analysis and results

5.1. Main result and discussion

We are finally in a position to state and prove our main result for the convergence rate of (OMD). Since we make no global monotonicity assumptions for the defining vector field v, this result is *de facto* of a local nature. We provide below the complete statement, followed by a series of remarks and an outline of the main steps of the proof.

Theorem 1. *Let x^* be a solution of (VI) and fix some confidence level $\delta > 0$. Suppose further that Assumptions 1–3 hold and (OMD) is run with a step-size of the form $\gamma_t = \gamma/(t + t_0)^\eta$ with $\eta \in (1/2, 1]$ and $\gamma, t_0 > 0$. We then have the following cases:*

Case I: *If the Legendre exponent of h at x^* is $\beta \in (0,1)$ and $\min\{\gamma, 1/t_0\}$ is small enough, then:*

(a) *x^* is stochastically stable: for every neighborhood $\mathcal{U}$ of x^*, there exists a neighborhood $\mathcal{U}_1$ of x^* such that the event*

$$\mathcal{E}_{\mathcal{U}} = \{X_t \in \mathcal{U} \text{ for all } t = 1, 2, \dots\} \tag{11}$$

occurs with probability at least $1 - \delta$ if $X_1 \in \mathcal{U}_1$.

(*b*) *If $\mathcal{U}$ is small enough and $X_1 \in \mathcal{U}_1$, the iterates X_t of* (OMD) *enjoy the convergence rate*

$$\mathbb{E}[D(x^*, X_t) \,|\, \mathcal{E}_{\mathcal{U}}] = \begin{cases} \mathcal{O}\big((\log t)^{-(1-\beta)/\beta}\big) & \text{if } \eta = 1, \\ \mathcal{O}\big(t^{-\min\{(1-\eta)(1-\beta)/\beta, \eta\}}\big) & \text{if } \frac{1}{2} < \eta < 1, \end{cases} \tag{12}$$

In particular, this rate undergoes a phase transition at $\beta = 1/2$:

- *Euclidean-like phase: If $\beta \in (0, 1/2)$ and $\eta = 1 - \beta$, we obtain the optimized rate*

$$\mathbb{E}[D(x^*, X_t) \,|\, \mathcal{E}_{\mathcal{U}}] = \mathcal{O}\Big(t^{-(1-\beta)}\Big). \tag{13}$$

- *Entropy-like phase: If $\beta \in [1/2, 1)$ and $\eta = (1+\epsilon)/2$ for any small $\varepsilon > 0$, we get*

$$\mathbb{E}[D(x^*, X_t) \,|\, \mathcal{E}_{\mathcal{U}}] = \mathcal{O}\Big(t^{-\frac{1-\varepsilon}{2\beta}(1-\beta)}\Big). \tag{14}$$

Case II: *If $\beta = 0$ and γ and t_0 are sufficiently large, x^* is stochastically stable and the iterates X_t of* (OMD) *enjoy the rate*

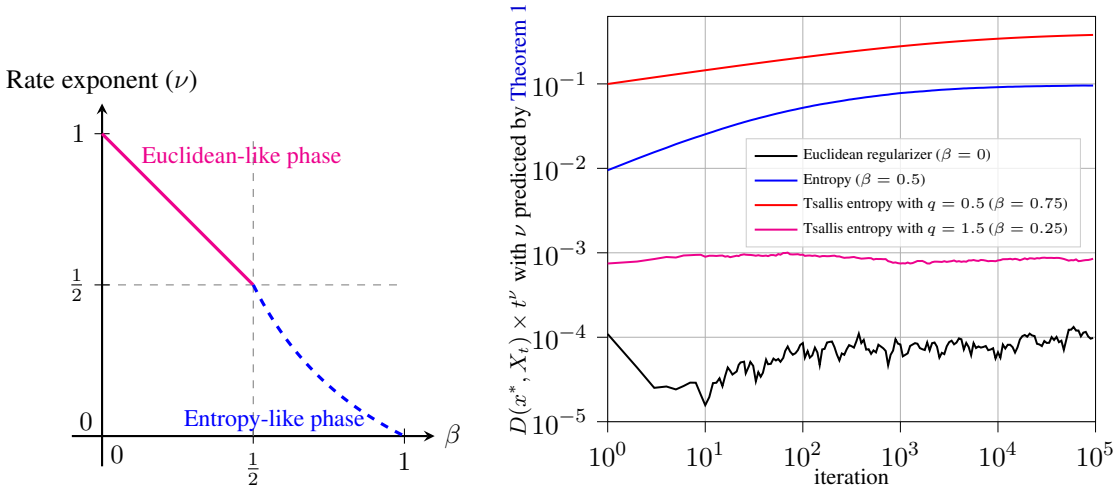
$$\mathbb{E}[D(x^*, X_t) \,|\, \mathcal{E}_{\mathcal{U}}] = \mathcal{O}(t^{-\eta}), \tag{15}$$

provided that $\mathcal{U}$ and $\mathcal{U}_1$ are small enough and $X_1 \in \mathcal{U}_1$. In particular, for $\eta = 1$ we get

$$\mathbb{E}[D(x^*, X_t) \,|\, \mathcal{E}_{\mathcal{U}}] = \mathcal{O}(t^{-1}). \tag{16}$$

Related work. The rate guarantees of Theorem 1 should be compared and contrasted to a series of recent results on the convergence speed of OMD and its variants. Most of these results concern the deterministic case, i.e., when $U_t = 0$ for all t. In this regime, and focusing on the last iterate of the method, Golowich et al. (2020a,b) showed that the convergence speed of the (optimistic) extra-gradient algorithm in unconstrained, smooth, monotone variational inequalities is $\mathcal{O}(1/\sqrt{t})$ in terms of $\|v(X_t)\|$. If, in addition, v is *strongly* monotone, it is well known that this rate becomes linear; for a recent take on different (Euclidean) variants of (OMD), see Malitsky (2015), Gidel et al. (2019), Hsieh et al. (2019), Mokhtari et al. (2019a,b), and references therein.

Unsurprisingly, the stochastic regime is fundamentally different. In the merely monotone case (i.e., neither strongly nor strictly monotone), we are not aware of any result on the method's last-iterate convergence speed: existing results in the literature either cover the algorithm's ergodic average (Cui and Shanbhag, 2016; Gidel et al., 2019; Juditsky et al., 2011) or are asymptotic in nature (Iusem et al., 2017). Because of the lack of deterministic correlation between gradients at the base and leading states in stochastic problems, smoothness does not help to achieve better rates in this case – except for mollifying the requirement for bounded gradient signals (Juditsky et al., 2011). By contrast, strong monotonicity – local or global – *does* help: as was shown by Hsieh et al. (2019), the iterates of the *Euclidean* version of (OMD) run with a $1/t$ step-size schedule converge locally to solutions satisfying (SOS) at a $\mathcal{O}(1/t)$ rate in terms of the mean square distance to the solution. This result is a special instance of Case II of Theorem 1: in particular, for $\beta = 0$, $\eta = 1$, (15) indicates that the $1/t$ step-size schedule examined by Hsieh et al. (2019) is optimal for the Euclidean case; however, as we explain below, it is *not* optimal for non-Euclidean DGFs.



(a) Predicted rate $\mathcal{O}(1/t^\nu)$ vs. Legendre exponent β

(b) Fit of the predicted rates with the simulation.

Figure 1: Illustration of the rates predicted by Theorem 1 (left: theory; right: numerical validation). The dashed line in Fig. 1a indicates that the rate prediction is only valid up to an arbitrarily small exponent or suitable logarithmic factor. The experiments in Fig. 1b were conducted for a linear vector field $v(x) = x$ on $\mathcal{K} = [0, +\infty)$ with the regularizer setups of Examples 3.1–3.3 (for more details, see Appendix D). We run OMD with the stepsizes prescribed by Theorem 1 and plot $D(x^*, X_t) \times t^\nu$ where ν is the predicted rate ($\nu = 1, 0.5, 0.167, 0.75$ respectively for Euclidean, entropy, Tsallis entropy with $q = 0.5$ and $q = 1.5$). A horizontal line corresponds to a match between the theoretical and observed rates.

Step-size tuning. A key take-away from Theorem 1 is that the step-size schedule that optimizes the rate guarantees in Eqs. (13) and (14) is either $\gamma_t = \Theta(1/t^{1-\beta})$ when $0 \leq \beta < 1/2$, or as close as possible to $\Theta(t^{1/2})$ when $1/2 \leq \beta < 1$; in both cases, the optimal tuning of the algorithm's step-size *depends on the Legendre exponent of* x^*. We feel that this is an important parameter to keep in mind when deploying a non-Euclidean version of (OMD) in a stochastic setting – e.g., as in the optimistic multiplicative weights update (OMWU) algorithm of Daskalakis and Panageas (2019). Theorem 1 suggests that the "best practices" of the Euclidean setting do not carry over to non-Euclidean ones: Euclidean step-size policy $1/t$ would lead to a catastrophic drop from $1/\sqrt{t}$ to $1/\log t$ in the rate guarantees for (OMD).

Geometry-dependent rates. One last observation for Theorem 1 is that the convergence speed is measured in terms of the Bregman divergence $D(x^*, X_t)$. Since $\|X_t - x^*\| = \mathcal{O}(\sqrt{D(x^*, X_t)})$, this rate immediately translates to a rate on $\|X_t - x^*\|$. However, this rate is *not* tight, again because of the disparity between the Bregman and norm geometries discussed in Section 4. Specifically, in many cases, a non-zero Legendre exponent implies a sharper lower bound on the Bregman divergence of the form $D(x^*, X_t) = \Omega(\|X_t - x^*\|^{2-2\beta})$. Thus, for $\beta \in [0, 1/2)$, Jensen's inequality applied to the optimized bound (13) gives $\mathbb{E}[\|X_t - x^*\|] = \mathcal{O}(1/\sqrt{t})$. Importantly, this rate guarantee is the same as in the Euclidean case ($\beta = 0$), though it requires a completely different step-size policy to achieve it. Determining when the Legendre exponent also provides a lower bound is a very interesting direction for future research, but one that lies beyond the scope of this work.

5.2. Main ideas of the proof

The heavy lifting in the proof of Theorem 1 is provided by Proposition 1 below, which also makes explicit some of the hidden constants in the statement of the theorem.

Proposition 1. *With notation and assumptions as in Theorem 1, fix some $r > 0$ and let*

$$\mathcal{U} = \mathbb{B}(x^*, r) \cap \mathcal{K} \quad \text{and} \quad \mathcal{U}_1 = \{x : D(x^*, x) \leq r^2/12\}.$$

If $X_1 \in \mathcal{U}_1$, then the event $\mathcal{E}_{\mathcal{U}}$ defined in (11) satisfies $\mathbb{P}(\mathcal{E}_{\mathcal{U}}) \geq 1 - \mathcal{O}\big(\rho^2/r^2\big)$ where $\rho^2 = \sigma^2\gamma^2 t_0^{1-2\eta}/(2\eta - 1)$. Moreover, if r is taken small enough, $\mathbb{E}[D(x^*, X_t) \mid \mathcal{E}_{\mathcal{U}}]$ is bounded according to the following table and conditions:

Legendre exponent	*Rate* ($\eta = 1$)	*Rate* ($\frac{1}{2} < \eta < 1$)
$\beta \in (0, 1)$	$\mathcal{O}\Big((\log t)^{-\frac{1-\beta}{\beta}}\Big)$	$\mathcal{O}\Big(t^{-\frac{(1-\eta)(1-\beta)}{\beta}} + t^{-\eta}\Big)$
Conditions:	$\gamma^{1+1/\beta} \leq \frac{c(\eta,\alpha)\max(1,\Phi)}{4\sigma^2(\mu/\kappa)^{1/\beta-1}}$	
$\beta = 0$	$\mathcal{O}(1/t)$	$\mathcal{O}(1/t^\eta)$
Conditions:	$\gamma > \kappa/\mu$	—

where $\alpha = \beta/(1-\beta)$, $\Phi = r^2/12 + 8\rho^2$, and

$$c(\eta, \alpha) = \begin{cases} \frac{1-\eta}{(1+\alpha)^{1+1/\alpha}} & \text{if } \eta \leq \frac{1+\alpha}{1+2\alpha}, \\ \alpha\left(\frac{1-2^{-(1+\alpha)}}{1+\alpha}\right)^{1+1/\alpha} & \text{if } \eta > \frac{1+\alpha}{1+2\alpha}. \end{cases} \tag{17}$$

We outline here the main ideas of the proof, deferring all details to Appendix C.

Sketch of proof. Our proof strategy is to show that, under the stated conditions for r and the algorithm's initialization, the iterates of (OMD) remain within a neighborhood of x^* where (SOS) holds with high probability. Then, conditioning on this event, we use the Bregman divergence as a stochastic potential function, and we derive the algorithm's rate of convergence using a series of lemmas on (random) numerical sequences. The step-by-step process is as follows:

1. The cornerstone of the proof is a descent inequality for the iterates of OMD, which relates $D(x^*, X_{t+1})$ to $D(x^*, X_t)$. As stated in Lemma B.2, we have

$$\begin{aligned} D(x^*, X_{t+1}) + \phi_{t+1} \leq D(x^*, X_t) + (1 - \gamma_t\mu)\phi_t - \gamma_t\langle V_{t+1/2}, X_{t+1/2} - x^*\rangle \\ + \left(4\gamma_t^2 L^2 - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2 + 4\gamma_t^2\left[\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2\right] \end{aligned} \tag{18}$$

   where $\phi_t = \frac{\gamma_{t-1}^2}{2}\|V_{t-1/2} - V_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = 0$.

2. The next step is to show that, with high probability and if initialized close to x^*, the iterates of OMD remain in $\mathcal{U}$. This is a technical argument relying on a use of the Doob-Kolmogorov maximal inequality for submartingales (in the spirit of Hsieh et al. (2020)) and the condition $\eta > 1/2$ (which in turn guarantees that the step-size is square summable, so the corresponding submartingale error terms are bounded in L^1). The formal statement is Lemma B.4.

3. To proceed, choose $r > 0$ small enough so that the second order sufficiency condition (SOS) holds and the Legendre exponent estimate (10) both hold. More precisely, we take r small enough so that $\mathcal{U} = \mathbb{B}(x^*, r) \cap \mathcal{K}$ is included in the domains of validity of both properties ($\mathcal{B}$ and $\mathcal{V}$ respectively), and we will work on the event $\mathcal{E}_t = \{X_{s+1/2} \in \mathcal{U} \text{ for all } s = 1, 2, \dots, t\}$. Then, conditioning on $\mathcal{E}_t$, Assumption 2 gives

$$\langle v(X_{t+1/2}), X_{t+1/2} - x^* \rangle \geq \frac{\mu}{2} \|X_{t+1/2} - x^*\|^2 ,$$

and hence, by Lemma A.6, we get

$$\langle v(X_{t+1/2}), X_{t+1/2} - x^* \rangle \geq \frac{\mu}{2} \|X_t - x^*\|^2 - \mu \|X_{t+1/2} - X_t\|^2 .$$

However, the descent inequality only involves $D(x^*, X_t)$ and not $\|X_t - x^*\|^2$, so we need to relate the latter to the former. This is where the Legendre exponent comes into play. More precisely, Definition 2 with $1 + \alpha = \frac{1}{1-\beta}$ gives

$$D(x^*, x)^{1+\alpha} \leq \tfrac{1}{2} \kappa \|x^* - x\|^2,$$

so, with $X_{t+1/2} \in \mathcal{V}$, we have

$$\langle v(X_{t+1/2}), X_{t+1/2} - x^* \rangle \geq \frac{\mu}{\kappa} D(x^*, X_t)^{1+\alpha} - \mu \|X_{t+1/2} - X_t\|^2 .$$

4. Thus, going back to the descent inequality (18), employing the bound derived above, and taking expectations ultimately yields (after some technical calculations) the iterative bound

$$\begin{aligned} &\mathbb{E}\left[(D(x^*, X_{t+1}) + \phi_{t+1}) \mathbb{1}_{\mathcal{E}_t}\right] \\ &\qquad \leq \mathbb{E}\left[\left(D(x^*, X_t) - \frac{\mu \gamma_t}{\kappa} D(x^*, X_t)^{1+\alpha} + \left(1 - \frac{\gamma_t \mu}{\kappa}\right) \phi_t\right) \mathbb{1}_{\mathcal{E}_{t-1}}\right] + 8 \gamma_t^2 \sigma^2 \end{aligned} \tag{19}$$

where we used Assumption 3 to bound the expectations of $\|U_{t+1/2}\|_*^2$ and $\|U_{t-1/2}\|_*^2$.

5. We are now in a position to reduce the problem under study to the behavior of the sequence $a_t = \mathbb{E}\left[(D(x^*, X_t) + \phi_t) \mathbb{1}_{\mathcal{E}_{t-1}}\right]$, $t \geq 1$. Indeed, taking a closer look at $\mathcal{E}_t$ shows that

$$\mathbb{E}[D(x^*, X_t) \mid \mathcal{E}_{\mathcal{U}}] = \mathcal{O}(a_t) .$$

Consequently, the inequality (19) above can be rewritten as follows:

$$a_{t+1} \leq a_t - \frac{\mu \gamma_t}{\kappa} \mathbb{E}\left[\left(D(x^*, X_t)^{1+\alpha} + \phi_t\right) \mathbb{1}_{\mathcal{E}_{t-1}}\right] + 8 \gamma_t^2 \sigma^2 . \tag{20}$$

6. The behavior of this last sequence hinges heavily on whether $\alpha > 0$ or not. In detail, we have:

   - If $\alpha = 0$, (20) gives
   $$a_{t+1} \leq \left(1 - \frac{\mu \gamma}{\kappa (t + t_0)^\eta}\right) a_t + \frac{8 \gamma^2 \sigma^2}{(t + t_0)^{2\eta}} .$$
   The long-run behavior of this recursive inequality is described by Lemmas A.8 and A.9: as long as $\mu \gamma / \beta > 1$ for $\eta = 1$ (and only for $\eta = 1$), we have $a_t = \mathcal{O}(1/t^\eta)$.

- If $\alpha > 0$, a series of further technical calculations in the same spirit yields

$$a_{t+1} \leq a_t - \frac{\mu\gamma}{2^\alpha \max(1, \Phi^\alpha)\kappa(t+t_0)^\eta} a_t^{1+\alpha} + \frac{8\gamma^2\sigma^2}{(t+t_0)^{2\eta}}$$

  The final step of our proof is to upper bound the behavior of a_t based on this recursive inequality. The necessary groundwork is provided by Lemmas A.11 and A.12, and hinges on whether $\eta \geq \frac{1+\alpha}{1+2\alpha}$ or not (and is also where the value of $c(\eta, \alpha)$ comes in).

Putting everything together, we obtain the conditional rates in the statement of the proposition. ■

6. Conclusion

In this work, we investigated the rate of convergence of optimistic mirror descent for solving variational inequalities using stochastic oracles. Our results highlight the relationship between the rate of convergence of the last iterate and the local geometry of the distance-generating function at the solution. To capture this local geometry, we introduced the regularity exponent of the divergence relative to the norm that we dubbed the Legendre exponent. This quantity plays a central role in our results: the less regular the Bregman divergence around the solution, the slower the convergence of the last iterate. Furthermore, we show that that the stepsize policy that guarantees the best rates depend on the distance-generating function through the associated Legendre exponent.

This work opens the door to various refinements and extensions diving deeper into the geometry of the method's DGF. A key remark is that the method's Legendre exponent seems to depend crucially on which constraints of the problem are active at a given solution. Deriving a precise characterization of the method's convergence rate in these cases is a fruitful direction for future research.

References

Felipe Alvarez, Jérôme Bolte, and Olivier Brahic. Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, 43(2):477–501, 2004.

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

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

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

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

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT '12: Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Kuo-Liang Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3): 463–483, 1954.

Shisheng Cui and Uday V. Shanbhag. On the analysis of reflected gradient and splitting methods for monotone stochastic variational inequality problems. In *CDC '16: Proceedings of the 57th IEEE Annual Conference on Decision and Control*, 2016.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS '19: Proceedings of the 10th Conference on Innovations in Theoretical Computer Science*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT '20: Proceedings of the 33rd Annual Conference on Learning Theory*, 2020b.

Elad Hazan. A survey: The convex optimization approach to regret minimization. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, pages 287–304. MIT Press, 2012.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6936–6946, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Alfredo N. Iusem, Alejandro Jofré, Roberto I. Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

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

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

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

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. https://arxiv.org/abs/1901.08511v2, 2019a.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. https://arxiv.org/pdf/1906.01115.pdf, 2019b.

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

Arkadi Semen Nemirovski and David Berkovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY, 1983.

Arkadi Semen Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

Boris Teodorovich Polyak. *Introduction to Optimization*. Optimization Software, New York, NY, USA, 1987.

Leonid Denisovich Popov. 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.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*, 2013a.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013b.

Ralph Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Optimization*, 14(5):877–898, 1976.

Constantino Tsallis. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52: 479–487, 1988.

John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. Translated by S. Bargmann as "On the Theory of Games of Strategy" in A. Tucker and R. D. Luce, editors, *Contributions to the Theory of Games IV*, volume 40 of *Annals of Mathematics Studies*, pages 13-42, 1957, Princeton University Press, Princeton.

Appendix A. Auxiliary lemmas

We gather in this section the basic results used in our proofs.

A.1. Bregman divergences

We recall here classical lemmas on Bregman divergences; see Juditsky et al. (2011); Mertikopoulos et al. (2019). We complement them with two elementary results Lemma A.6 on a specific bounding and Lemma A.7 which shows that the Legendre exponent is typically trivial on $\mathcal{K}_h$.

Lemma A.1. *For $x \in \mathcal{K}$, $p \in \operatorname{dom} \partial h$,*

$$D(x,p) \geq \frac{1}{2}\|p-x\|^2 .$$

Lemma A.2. *For $x \in \operatorname{dom} \partial h$, $y, y' \in \mathcal{Y}$,*

$$\|P_x(y) - P_x(y')\| \leq \|y - y'\|_* .$$

Lemma A.3. *For $p \in \mathcal{K}$, $x \in \operatorname{dom} \partial h$, $y \in \mathcal{Y}$, $y' \in N_{\mathcal{K}}(p)$ and $x^+ = P_x(y)$,*

$$\begin{aligned} D(p, x^+) &\leq D(p, x) + \langle y - y', x^+ - p \rangle - D(x^+, x) \\ &\leq D(p, x) + \langle y - y', x - p \rangle + \frac{1}{2}\|y - y'\|_*^2 . \end{aligned}$$

Lemma A.4. *For $p \in \mathcal{K}$, $x \in \operatorname{dom} \partial h$, $y, y' \in \mathcal{Y}$ and $x_1^+ = P_x(y)$,$x_2^+ = P_x(y')$,*

$$D(p, x_2^+) \leq D(p, x) + \langle y', x_1^+ - p \rangle + \frac{1}{2}\|y' - y\|_*^2 - \frac{1}{2}\|x_1^+ - x\|^2 .$$

Lemma A.5. *For $x \in \mathcal{K}_h$,*

$$\partial h(x) = \nabla h(x) + \mathrm{NC}_{\mathcal{K}}(x) .$$

As a consequence, $x \in \mathcal{K}_h$, $x^+ \in \mathcal{K}$, $y \in \mathcal{Y}$,

$$\begin{aligned} x^+ = P_x(y) &\iff \nabla h(x) + y \in \partial h(x^+) \\ &\iff \nabla h(x) + y - \nabla h(x^+) \in \mathrm{NC}_{\mathcal{K}}(x^+) . \end{aligned}$$

Moreover, $x^+ = P_x(y)$ implies that $x^+ \in \mathcal{K}_h h$.

Lemma A.6. *If Assumption 2 holds, $x^+ \in \mathcal{B}$, and $x \in \mathcal{K}$, then,*

$$\langle v(x^+), x^+ - x^* \rangle \geq \frac{\mu}{2}\|x - x^*\|^2 - \mu\|x^+ - x\|^2 .$$

Proof. If $x^+ \in \mathcal{B}$, then,

$$\langle v(x^+), x^+ - x^* \rangle \geq \mu\|x^+ - x^*\|^2 . \tag{A.1}$$

However, we are interested in the distance between x and x^*, not in the distance between x^+ and x^*. To remedy this, we use Young's inequality,

$$\|x - x^*\|^2 \leq 2\|x - x^+\|^2 + 2\|x^+ - x^*\|^2 ,$$

which leads to,

$$\|x^+ - x^*\|^2 \geq \frac{1}{2}\|x - x^*\|^2 - \|x^+ - x\|^2 .$$

Combined with Eq. (A.1), we get,

$$\langle v(x^+), x^+ - x^* \rangle \geq \frac{\mu}{2}\|x - x^*\|^2 - \mu\|x^+ - x\|^2 .$$

■

Lemma A.7 (Trivial Legendre exponent on the interior)**.** *Assume that ∇h is locally Lipschitz continuous on $\mathcal{K}_h$. Then, for any $p \in \mathcal{K}_h$, the Legendre exponent of h at p is $\beta = 0$.*

Proof. Take $p \in \mathcal{K}_h$. As ∇h is assumed to be locally Lipschitz, there exists $\mathcal{V}$ a neighborhood of p and $\kappa > 0$ such that, for any $x \in \mathcal{V} \cap \mathcal{K}_h$,

$$\|\nabla h(p) - \nabla h(x)\|_* \leq \frac{\kappa}{2}\|p - x\| .$$

Now, as $\nabla h(p) \in \partial h(p)$, for any $x \in \mathcal{V} \cap \mathcal{K}_h$,

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

Using the local Lipschitz continuity of ∇h, we get, for any $x \in \mathcal{V} \cap \mathcal{K}_h$,

$$D(p,x) \leq \frac{\kappa}{2} \|p - x\|^2 \,,$$

which reads $\beta = 0$. ■

A.2. Sequences

We recall results on sequences to transform descent-like inequalities into rates; see the classical textbook Polyak (1987). We also need variants of existing results (Lemma A.11 and Lemma A.12) that we state here.

Lemma A.8 (Chung (1954, Lem. 1))**.** *Let $(a_t)_{t \geq 1}$ be a sequence of non-negative scalars, $q > 1$, $q' > 0$, $t_0 \geq 0$. If, for any $t \geq 1$,*

$$a_{t+1} \leq \left(1 - \frac{q}{t + t_0}\right) a_t + \frac{q'}{(t + t_0)^2} \,,$$

then, for any $T \geq 1$,

$$a_T \leq \frac{q'}{q - 1} \frac{1}{T + t_0} + o\left(\frac{1}{T}\right) \,.$$

Lemma A.9 (Chung (1954, Lem. 4))**.** *Let $(a_t)_{t \geq 1}$ be a sequence of non-negative scalars, $q, q' > 0$, $t_0 \geq 0$, $1 > \eta > 0$. If, for any $t \geq 1$,*

$$a_{t+1} \leq \left(1 - \frac{q}{(t + t_0)^\eta}\right) a_t + \frac{q'}{(t + t_0)^{2\eta}} \,,$$

then, for any $T \geq 1$,

$$a_T \leq \mathcal{O}\left(\frac{1}{T^\eta}\right) \,.$$

Lemma A.10 (Polyak (1987, §2.2, Lem. 6))**.** *Let $(a_t)_{1 \leq t \leq T}$ be a sequence of non-negative scalars, $(\gamma_t)_{1 \leq t \leq T}$ be sequence of positive scalars and $\alpha > 0$. If, for any $t = 1, \ldots, T$,*

$$a_{t+1} \leq a_t - \gamma_t a_t^{1+\alpha} \,,$$

then,

$$a_T \leq \frac{a_1}{\left(1 + \alpha a_1^\alpha \sum_{t=1}^{T-1} \gamma_t\right)^{1/\alpha}} \,.$$

Though the next two lemmas allow for step-sizes of the form $\gamma_t = \frac{\gamma}{(t+t_0)^\eta}$, they do not exploit the term t_0.

Lemma A.11. *Let $(a_t)_{t\geq 1}$ be a sequence of non-negative scalars, $(\gamma_t)_{t\geq 1}$ be sequence of positive scalars of the form $\gamma_t = \frac{q}{(t+t_0)^\eta}$ with $q, \eta > 0$, $t_0 \geq 0$ and $\alpha > 0$, $q' > 0$ such that,*

$$a_{t+1} \leq a_t - \gamma_t a_t^{1+\alpha} + \frac{q'}{(t+t_0)^{2\eta}}\,.$$

If,

$$1 \geq \eta \geq \frac{1+\alpha}{1+2\alpha} > \frac{1}{2}\,, \qquad q' q^{1/\alpha} \leq c(\eta,\alpha) \coloneqq \alpha\left(\frac{1-2^{-(1+\alpha)}}{1+\alpha}\right)^{1+\frac{1}{\alpha}}\,,$$

then, for any $T \geq 1$,

$$a_T \leq \frac{a_1 + b}{\left(1 + \alpha(a_1+b)^\alpha 2^{-\alpha}\sum_{t=1}^{T-1}\gamma_t\right)^{1/\alpha}}\,,$$

where $b = \left(\frac{1-2^{1-2\eta}}{(1+\alpha)q}\right)^{\frac{1}{\alpha}}$.

Proof. First, define $p = 1+\alpha > 1$ and note that $\frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1} = \frac{1}{2-p^{-1}} \in (\frac{1}{2}, 1)$ so, in particular, the condition on η is not absurd. Moreover, this means that $\beta \coloneqq 2\eta - 1$ belongs to $(0,1]$. Then, we have $b = \left(\frac{1-2^{-\beta}}{pq}\right)^{\frac{1}{\alpha}} > 0$ and, define, for $t \geq 1$, $b_t \coloneqq \frac{b}{(t+t_0)^\beta}$.

The first part of the proof consists in showing, that, for $t \geq 1$,

$$\frac{q'}{(t+t_0)^{2\eta}} \leq b_t - b_{t+1} - \gamma_t b_t^p\,.$$

For this, we will need the following remark. As $\beta \leq 1$, $x \mapsto (1+x)^\beta$ is concave. Hence, it is above its chords, and in particular above its chord going from 0 to 1. Thus, for $0 \leq x \leq 1$, $(1+x)^\beta \geq 1 + x(2^\beta - 1)$.

We use this remark to lower bound $b_t - b_{t+1}$ for $t \geq 1$. Indeed,

$$\begin{aligned}
b_t - b_{t+1} &= \frac{b}{(t+t_0)^\beta} - \frac{b}{(t+1+t_0)^\beta} \\
&= \frac{b}{(t+1+t_0)^\beta}\left(\left(1 + \frac{1}{t+t_0}\right)^\beta - 1\right) \\
&\geq \frac{b}{(t+1+t_0)^\beta}\frac{2^\beta - 1}{t+t_0} \\
&\geq \frac{b}{(t+t_0)^{\beta+1}}\frac{2^\beta - 1}{2^\beta}\,.
\end{aligned}$$

Therefore,

$$b_t - b_{t+1} - \gamma_t b_t^p \geq \frac{b}{(t+t_0)^{\beta+1}}(1 - 2^{-\beta}) - \frac{q b^p}{(t+t_0)^{\eta+p\beta}}\,.$$

Now, by the definition of β, $\beta + 1 = 2\eta$ and $\eta + p\beta = (2p+1)\eta - p = 2\eta + (2p-1)\eta - p \geq 2\eta$ by the assumption that $\eta \geq \frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1}$. Hence,

$$b_t - b_{t+1} - \gamma_t b_t^p \geq \frac{1}{(t+t_0)^{2\eta}}(b(1-2^{-\beta}) - qb^p)\,,$$

so that we only need to show that $q' \leq b(1-2^{-\beta}) - qb^p$.

Rearranging and replacing b by its expression gives,

$$\begin{aligned} b(1-2^{-\beta}) - qb^p &= b((1-2^{-\beta}) - qb^\alpha) \\ &= b\left((1-2^{-\beta}) - \frac{1-2^{-\beta}}{p}\right) \\ &= b(1-2^{-\beta})\frac{p-1}{p} \\ &= (1-2^{-\beta})^{1+\frac{1}{\alpha}}\frac{p-1}{p^{1+\frac{1}{\alpha}}q^{\frac{1}{\alpha}}}\,. \end{aligned}$$

Therefore, with $c(\eta,\alpha) = \alpha\left(\frac{1-2^{-(1+\alpha)}}{1+\alpha}\right)^{1+\frac{1}{\alpha}} = (1-2^{-\beta})^{1+\frac{1}{\alpha}}\frac{p-1}{p^{1+\frac{1}{\alpha}}} > 0$ and $q^{\frac{1}{\alpha}}q' \leq c(\eta,\alpha)$, we finally get that $q' \leq b(1-2^{-\beta}) - qb^p$ and, for $t \geq 1$,

$$\frac{q'}{(t+t_0)^{2\eta}} \leq b_t - b_{t+1} - \gamma_t b_t^p\,.$$

Recall that $(a_t)_{t\geq 1}$ satisfies, for $t \geq 1$,

$$a_{t+1} \leq a_t - \gamma_t a_t^p + \frac{q'}{(t+t_0)^{2\eta}}\,.$$

Therefore, putting these two inequalities together gives,

$$a_{t+1} + b_{t+1} \leq a_t + b_t - \gamma_t(a_t^p + b_t^p)\,.$$

Finally, by convexity of $x \mapsto x^p$,

$$a_{t+1} + b_{t+1} \leq a_t + b_t - \gamma_t 2^{1-p}(a_t + b_t)^p\,.$$

Now, we can apply Lemma A.10 with $a_t \leftarrow a_t + b_t$ and $\gamma_t \leftarrow \gamma_t 2^{1-p}$. For any $T \geq 1$,

$$a_T \leq a_T + b_T \leq \frac{1}{\left((a_1+b_1)^{-\alpha} + \alpha 2^{-\alpha}\sum_{t=1}^{T-1}\gamma_t\right)^{1/\alpha}}\,.$$

As the right-hand side is non-decreasing in b_1 and $b_1 \leq b$, we get the result of the statement. ■

Lemma A.12. *Let $(a_t)_{t\geq 1}$ be a sequence of non-negative scalars, $(\gamma_t)_{t\geq 1}$ be sequence of positive scalars of the form $\gamma_t = \frac{q}{(t+t_0)^\eta}$ with $q, \eta > 0$, $b \geq 1$ and $\alpha > 0$, $q' > 0$ such that,*

$$a_{t+1} \leq a_t - \gamma_t a_t^{1+\alpha} + \frac{q'}{(t+t_0)^{2\eta}}\,.$$

If,

$$1 > \frac{1+\alpha}{1+2\alpha} \geq \eta > \frac{1}{2}\,, \qquad\qquad q' q^{1/\alpha} \leq c(\eta,\alpha) \coloneqq \frac{1-\eta}{(1+\alpha)^{\frac{1}{\alpha}+1}}\,,$$

then, for any $T \geq 1$*,*

$$a_T \leq \frac{a_1}{\left(1 + \alpha a_1^\alpha \sum_{t=1}^{T-1} \gamma_t\right)^{1/\alpha}} + \frac{1}{((1+\alpha)q)^{\frac{1}{\alpha}}(T+t_0)^\eta}\,.$$

Proof. This proof is the "mirror" of the proof of the previous lemma. As before, define $p = 1+\alpha > 1$ and note that $\frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1} = \frac{1}{2-p^{-1}} \in (\frac{1}{2}, 1)$ so the condition on η is not absurd. Moreover, this means that $\beta \coloneqq \frac{\eta}{p}$ belongs to $(0,1)$. Then, define $b \coloneqq \left(\frac{1}{pq}\right)^{\frac{1}{\alpha}} > 0$ and, define, for $t \geq 1$, $b_t \coloneqq \frac{b}{(t+t_0)^\beta}$.

Opposite to the proof of the previous lemma, the first part of the proof consists in showing, that, for $t \geq 1$,

$$\frac{q'}{(t+t_0)^{2\eta}} \leq b_{t+1} + \gamma_t b_t^p - b_t\,.$$

For this, we use the concavity of $x \mapsto (1+x)^\beta$, as $\beta \leq 1$, so that, for $x \geq 0$, $(1+x)^\beta \leq 1 + \beta x$.

This remark enables us to upper bound $b_t - b_{t+1}$ for $t \geq 1$. Indeed,

$$\begin{aligned} b_t - b_{t+1} &= \frac{b}{(t+t_0)^\beta} - \frac{b}{(t+1+t_0)^\beta} \\ &= \frac{b}{(t+1+t_0)^\beta}\left(\left(1 + \frac{1}{t+t_0}\right)^\beta - 1\right) \\ &\leq \frac{b}{(t+1+t_0)^\beta}\frac{\beta}{t+t_0} \\ &\leq \frac{\beta b}{(t+t_0)^{\beta+1}}\,. \end{aligned}$$

Therefore,

$$b_t - b_{t+1} - \gamma_t b_t^p \leq \frac{\beta b}{(t+t_0)^{\beta+1}} - \frac{q b^p}{(t+t_0)^{\eta+p\beta}}\,.$$

Now, by the definition of $\beta = \frac{\eta}{p}$, $\eta + p\beta = 2\eta$ and $\beta + 1 = 2\eta + \frac{\eta}{p} + 1 - 2\eta = 2\eta + \frac{p-(2p-1)\eta}{p} \geq 2\eta$ by the assumption that $\eta \leq \frac{1+\alpha}{1+2\alpha} = \frac{p}{2p-1}$. Hence,

$$b_t - b_{t+1} - \gamma_t b_t^p \leq \frac{1}{(t+t_0)^{2\eta}}(\beta b - q b^p)\,,$$

so that,

$$b_{t+1} + \gamma_t b_t^p - b_t \geq \frac{1}{(t+t_0)^{2\eta}}(q b^p - \beta b)\,.$$

Again, we only need to show that $q' \leq q b^p - \beta b$.

Rearranging and replacing b by its expression gives,

$$\begin{aligned} qb^p - b\beta &= b(qb^\alpha - \beta) \\ &= b\left(\frac{1}{p} - \beta\right) \\ &= \frac{1}{q^{\frac{1}{\alpha}} p^{\frac{1}{\alpha}}}\left(\frac{1}{p} - \beta\right). \end{aligned}$$

Therefore, with $c(\eta, \alpha) = \frac{1}{p^{\frac{1}{\alpha}}}\left(\frac{1}{p} - \beta\right) > 0$ and $q^{\frac{1}{\alpha}} q' \leq c(\eta, \alpha)$, we finally get that $q' \leq qb^p - \beta b$ and, for $t \geq 1$,

$$\frac{q'}{(t+t_0)^{2\eta}} \leq b_{t+1} + \gamma_t b_t^p - b_t\,.$$

Putting this inequality together with the one on the sequence $(a_t)_{t\geq 1}$ gives,

$$a_{t+1} - b_{t+1} \leq a_t - b_t - \gamma_t(a_t^p - b_t^p)\,.$$

Now, let us discuss separately the case when $a_t > b_t$ and when $a_t \leq b_t$.

More precisely, define $T_0 = \min\{t \geq 1 : a_t \leq b_t\} \in \mathbb{N}^* \cup \{+\infty\}$, so that, for any $1 \leq t < T_0$, $a_t > b_t > 0$. Note that, for $x, y > 0$, as $p > 1$, $x^p + y^p \leq (x+y)^p$. For $1 \leq t < T_0$, apply this with $x \leftarrow a_t - b_t$, $y \leftarrow b_t$ gives $(a_t - b_t)^p \leq a_t^p - b_t^p$ so that,

$$a_{t+1} - b_{t+1} \leq a_t - b_t - \gamma_t(a_t - b_t)^p\,.$$

Lemma A.10 with $a_t \leftarrow a_t - b_t$ gives, for $1 \leq T < T_0$,

$$\begin{aligned} a_T &\leq b_T + \frac{1}{((a_1 - b_1)^{-\alpha} + \alpha \sum_{t=1}^{T-1} \gamma_t)^{\frac{1}{\alpha}}} \\ &\leq b_T + \frac{1}{(a_1^{-\alpha} + \alpha \sum_{t=1}^{T-1} \gamma_t)^{\frac{1}{\alpha}}}\,, \end{aligned}$$

so the statement holds in this case.

We now handle the case of $T \geq T_0$. At $t = T_0$, we have $a_{T_0} \leq b_{T_0}$. We show, by induction, that, for any $t \geq T_0$, $a_t \leq b_t$. The initialization at $t = T_0$ is trivial. So, now, assume that $a_t \leq b_t$ for some $t \geq T_0$. Recall that we have,

$$a_{t+1} - b_{t+1} \leq a_t - b_t - \gamma_t(a_t^p - b_t^p)\,,$$

so we only need to show that,

$$a_t - \gamma_t a_t^p \leq b_t - \gamma_t b_t^p\,.$$

Define $\varphi_t : x \in \mathbb{R}_+ \mapsto x - \gamma_t x^p$. Differentiating this function shows that it is non-decreasing on $\{x \in \mathbb{R}_+ : p\gamma_t x^\alpha \leq 1\} = [0, (\gamma_t p)^{-\frac{1}{\alpha}}]$. But $a_t \leq b_t \leq b_1 \leq b$ by definition of $(b_s)_{s\geq 1}$ and, by definition of b and $(\gamma_s)_{s\geq 1}$, $b = (pq)^{-\frac{1}{\alpha}} = (\gamma_1 p)^{-\frac{1}{\alpha}} \leq (\gamma_t p)^{-\frac{1}{\alpha}}$.

Hence, both a_t and b_t belong to the interval on which φ_t is non-decreasing so that $\varphi_t(a_t) \leq \varphi_t(b_t)$, which implies that $a_{t+1} \leq b_{t+1}$. Therefore, we have shown that, for all $T \geq T_0$, $a_T \leq b_T = \frac{1}{(pq)^{\frac{1}{\alpha}}(T+t_0)^\eta}$ which concludes the proof. ■

Appendix B. Descent and stability of optimistic mirror descent

B.1. Descent inequality

The lemmas we present here link the divergence between two consecutive iterates and the solution.

In particular, the first one only rely on i) the definition of one iteration of optimistic mirror descent; and ii) the core properties of the Bregman regularizer i.e., the 1-strong convexity of h (through Lemma A.1) and the non-expansivity of the proximal mapping (Lemma A.2). It does not require any assumption on the vector field v, on the noise in V, or additional properties of the Bregman divergence (such as a Legendre exponent).

Lemma B.1. *If the stepsizes verify $\mu\gamma_t + 4L^2\gamma_t^2 \leq 1$, then, the iterates of optimistic mirror descent initialized with $X_{1/2} \in \mathcal{K}$, $X_1 \in \operatorname{dom} \partial h$ verify for all $t \geq 1$,*

$$\begin{aligned} D(x^*, X_{t+1}) + \phi_{t+1} \leq D(x^*, X_t) + (1 - \gamma_t\mu)\phi_t \\ - \gamma_t\langle V_{t+1/2}, X_{t+1/2} - x^*\rangle + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2 \\ + \gamma_t^2\Delta_{t+1}^2 - \gamma_t^2\tau^2\|X_t - X_{t-1/2}\|^2 \,, \end{aligned}$$

where $\Delta_{t+1}^2 = (\|V_{t+1/2} - V_{t-1/2}\|_*^2 - L^2\|X_{t+1/2} - X_{t-1/2}\|^2)_+$, $\phi_t = \frac{\gamma_{t-1}^2}{2}\|V_{t-1/2} - V_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2}\|X_1 - X_{1/2}\|^2$.

Proof. First, apply Lemma A.4 with $(x, y_1, y_2) \leftarrow (X_t, -\gamma_t V_{t-1/2}, -\gamma_t V_{t+1/2})$ and $p \leftarrow x^*$,

$$D(x^*, X_{t+1}) \leq D(x^*, X_t) - \gamma_t\langle V_{t+1/2}, X_{t+1/2} - x^*\rangle + \frac{\gamma_t^2}{2}\|V_{t-1/2} - V_{t+1/2}\| - \frac{1}{2}\|X_{t+1/2} - X_t\|^2 \,. \tag{B.1}$$

In order to upper-bound $\frac{\gamma_t^2}{2}\|V_{t-1/2} - V_{t+1/2}\| - \frac{1}{2}\|X_{t+1/2} - X_t\|^2$, we use the definition of $\phi_{t+1} = \frac{\gamma_t^2}{2}\|V_{t+1/2} - V_{t-1/2}\|^2$ to see that

$$\begin{aligned} \frac{\gamma_t^2}{2}\|V_{t-1/2} - V_{t+1/2}\|^2 &= \gamma_t^2\|V_{t-1/2} - V_{t+1/2}\|^2 - \frac{\gamma_t^2}{2}\|V_{t-1/2} - V_{t+1/2}\|^2 \\ &= \gamma_t^2\|V_{t-1/2} - V_{t+1/2}\|^2 - \phi_{t+1} \,, \end{aligned}$$

and the definition of Δ_{t+1} to bound

$$\begin{aligned} \gamma_t^2\|V_{t-1/2} - V_{t+1/2}\|^2 &\leq \gamma_t^2 L^2\|X_{t-1/2} - X_{t+1/2}\|^2 + \gamma_t^2\Delta_{t+1}^2 \\ &\leq 2\gamma_t^2 L^2\|X_{t-1/2} - X_t\|^2 + 2\gamma_t^2 L^2\|X_{t+1/2} - X_t\|^2 + \gamma_t^2\Delta_{t+1}^2 \,, \end{aligned}$$

where we used Young's inequality to get the last inequality line. Putting the two equations together, we get that

$$\begin{aligned} &\frac{\gamma_t^2}{2}\|V_{t-1/2} - V_{t+1/2}\|^2 - \frac{1}{2}\|X_{t+1/2} - X_t\|^2 \\ \leq &2\gamma_t^2 L^2\|X_{t-1/2} - X_t\|^2 + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2 - \phi_{t+1} + \gamma_t^2\Delta_{t+1}^2. \end{aligned} \tag{B.2}$$

For $t \geq 2$, we use the non-expansivity of the proximal mapping (Lemma A.2) to bound $\|X_{t-1/2} - X_t\|$ by $\gamma_{t-1}\|V_{t-3/2} - V_{t-1/2}\|_*$ in (B.2) to get

$$\begin{aligned}
&\frac{\gamma_t^2}{2}\|V_{t-1/2} - V_{t+1/2}\|^2 - \frac{1}{2}\|X_{t+1/2} - X_t\|^2 \\
&\leq (2\gamma_t^2 L^2)\gamma_{t-1}^2\|V_{t-3/2} - V_{t-1/2}\|_*^2 + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2 - \phi_{t+1} + \gamma_t^2\Delta_{t+1}^2 \\
&= (4\gamma_t^2 L^2)\phi_t + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2 - \phi_{t+1} + \gamma_t^2\Delta_{t+1}^2 \\
&\leq (1 - \mu\gamma_t)\,\phi_t + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2 - \phi_{t+1} + \gamma_t^2\Delta_{t+1}^2\,,
\end{aligned}$$

where we used the definition of ϕ_t and the assumption on the stepsizes. Combining this result with Eq. (B.1) gives the first assertion of the lemma for $t \geq 2$.

Now, if $t = 1$, Eq. (B.2) can be rewritten as,

$$\begin{aligned}
&\frac{\gamma_1^2}{2}\|V_{1/2} - V_{3/2}\|^2 - \frac{1}{2}\|X_{3/2} - X_1\|^2 \\
\leq&2\gamma_1^2 L^2\|X_{1/2} - X_1\|^2 + \left(2\gamma_1^2 L^2 - \frac{1}{2}\right)\|X_{3/2} - X_1\|^2 - \phi_2 + \gamma_1^2\Delta_2^2 \\
=&4\gamma_1^2 L^2\phi_1 + \left(2\gamma_1^2 L^2 - \frac{1}{2}\right)\|X_{3/2} - X_1\|^2 - \phi_2 + \gamma_1^2\Delta_2^2\,,
\end{aligned}$$

which yields the assertion of the lemma for $t = 1$ as $4\gamma_1^2 L^2 \leq 1 - \gamma_1\mu$. ■

We now use the properties of V and v to refine the bound of Lemma B.1 into a descent inequality.

Lemma B.2. *Let Assumptions 1 and 2 hold. Consider the iterates of optimistic mirror descent with:*

i) an unbiased stochastic oracle $V_t = v(X_t) + U_t$ (see (1));

ii) step-sizes $0 < \gamma_t \leq \frac{1}{4L}$.

Then,

$$\begin{aligned}
D(x^*, X_{t+1}) + \phi_{t+1} \leq& D(x^*, X_t) + (1 - \gamma_t\mu)\phi_t - \gamma_t\langle V_{t+1/2}, X_{t+1/2} - x^*\rangle \\
&+ \left(4\gamma_t^2 L^2 - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2 + 4\gamma_t^2(\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2)
\end{aligned} \tag{B.3}$$

where $\phi_t = \frac{\gamma_{t-1}^2}{2}\|V_{t-1/2} - V_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2}\|X_1 - X_{1/2}\|^2$.

Proof. First, let us examine the choice of γ_t. The condition $\gamma_t \leq \frac{1}{4L}$ is actually equivalent to,

$$8\gamma_t^2 L^2 + 2\gamma_t L \leq 1\,.$$

As $\mu \leq L$, this implies that,

$$8\gamma_t^2 L^2 + 2\gamma_t\mu \leq 1\,. \tag{B.4}$$

We first use the inequality of OMD provided by Lemma B.1 with $(\mu, L) \leftarrow (\mu, \sqrt{2}L)$ (note that its assumption is satisfied than4s to Eq. (B.4)) to get

$$\begin{aligned} D(x^*, X_{t+1}) + \phi_{t+1} \leq D(x^*, X_t) + (1 - \gamma_t\mu)\phi_t \\ - \gamma_t \langle V_{t+1/2}, X_{t+1/2} - x^* \rangle + \left(2\gamma_t^2 L^2 - \frac{1}{2}\right) \|X_{t+1/2} - X_t\|^2 \\ + \gamma_t^2 \Delta_{t+1}^2 - \gamma_t^2 \tau^2 \|X_t - X_{t-1/2}\|^2 \,, \end{aligned} \tag{B.5}$$

where $\Delta_{t+1}^2 = (\|V_{t+1/2} - V_{t-1/2}\|_*^2 - L^2\|X_{t+1/2} - X_{t-1/2}\|^2)_+$, $\phi_t = \frac{\gamma_{t-1}^2}{2}\|V_{t-1/2} - V_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2}\|X_1 - X_{1/2}\|^2$.

We now have to bound $\Delta_{t+1}^2 = (\|V_{t+1/2} - V_{t-1/2}\|_*^2 - 2L^2\|X_{t+1/2} - X_{t-1/2}\|^2)_+$ for $t \geq 1$. By using twice Young's inequality,

$$\begin{aligned} \|V_{t+1/2} - V_{t-1/2}\|_*^2 &\leq 2\|v(X_{t+1/2}) - v(X_{t-1/2})\|_*^2 + 2\|U_{t+1/2} - U_{t-1/2}\|_*^2 \\ &\leq 2\|v(X_{t+1/2}) - v(X_{t-1/2})\|_*^2 + 4\|U_{t+1/2}\|_*^2 + 4\|U_{t-1/2}\|_*^2 \,. \end{aligned}$$

Using the Lipschitz continuity of v (Assumption 1), we obtain that

$$\|V_{t+1/2} - V_{t-1/2}\|_*^2 \leq 2L^2\|X_{t+1/2} - X_{t-1/2}\|_*^2 + 4\|U_{t+1/2}\|_*^2 + 4\|U_{t-1/2}\|_*^2 \,,$$

so that,

$$\Delta_{t+1}^2 \leq 4\|U_{t+1/2}\|_*^2 + 4\|U_{t-1/2}\|_*^2 \,.$$

Plugging this inequality in (B.5), we obtain the claimed result. ■

B.2. Stochastic local stability

In order to use the local properties of the Bregman divergence around the solution, we need a local stability result that gives us that sufficiently close iterates will remain in a neighborhood of the solution with high probability. To show such a result, we will use the following lemma Hsieh et al. (2020, Lem. F.1) for bounding a recursive stochastic process.

Lemma B.3 (Hsieh et al. (2020, Lem. F.1))**.** *Consider a filtration $(\mathcal{F}_t)_t$ and four $(\mathcal{F}_t)_t$-adapted processes $(D_t)_t$, $(\zeta_t)_t$, $(\chi_t)_t$, $(\xi_t)_t$ such that χ_t is non-negative and the following recursive inequality is satisfied for all $t \geq 1$*

$$D_{t+1} \leq D_t - \zeta_t + \xi_t.$$

For $C > 0$, we define the events A_t by $A_1 \coloneqq \{D_1 \leq C/2\}$ and $A_t \coloneqq \{D_t \leq C\} \bigcap \{\chi_t \leq C/4\}$ for $t \geq 2$. We consider also the decreasing sequence of events I_t defined by $I_t \coloneqq \bigcap_{1 \leq s \leq t} A_s$. If the following three assumptions hold true

(i) $\forall t, \zeta_t \mathbb{1}_{I_t} \geq 0$,

(ii) $\forall t, \mathbb{E}[\xi_t \,|\, \mathcal{F}_t] \mathbb{1}_{I_t} = 0$,

(iii) $\sum_{t=1}^{\infty} \mathbb{E}[(\xi_{t+1}^2 + \chi_{t+1}) \mathbb{1}_{I_t}] \leq \delta\epsilon\, \mathbb{P}(A_1)$,

where $\epsilon = \min(C^2/16, C/4)$ and $\delta \in (0, 1)$, then $\mathbb{P}\left(\bigcap_{t \geq 1} A_t \mid A_1\right) \geq 1 - \delta$.

Lemma B.4. *Let Assumptions 1 and 2 hold. Consider the iterates of optimistic mirror descent with:*

i) an unbiased stochastic oracle $V_t = v(X_t) + U_t$ *with finite variance* σ^2 *(see (1) and (2));*

ii) step-sizes $0 < \gamma_t \leq \frac{1}{4L}$ *with* $\sum_{t=1}^{\infty} \gamma_t^2 = \Gamma < +\infty$.

Denote by $\mathcal{V}$ *the neighborhood of* x^* *on which Eq. (10) holds with* $p \leftarrow x^*$*. Then, for any* $r > 0$ *such that* $\mathcal{U} \coloneqq \mathbb{B}(x^*, r) \cap \mathcal{K}$ *is included in both* $\mathcal{V}$ *and* $\mathcal{B}$*,*

$$\mathbb{P}\left[\forall t \geq \frac{1}{2},\ X_t \in \mathcal{U} \middle| D(x^*, X_1) + \phi_1 \leq \frac{2r^2}{9}\right] \geq 1 - \frac{9(8 + r^2)\sigma^2\Gamma}{r^2 \min(1, r^2/9)}\,,$$

where $\phi_1 = \frac{1}{2}\|X_1 - X_{1/2}\|^2$.

Proof. This proof mainly consists in applying Lemma B.3 indexed by $t \in \frac{1}{2}\mathbb{N}^*,\ t \geq 1$. Define $C = \frac{r^2}{2+1/4}$ and, for $t \geq 1$, the following adapted processes,

$$\begin{aligned}
\zeta_t &\coloneqq 0 & \zeta_{t+1/2} &\coloneqq \gamma_t \langle v(X_{t+1/2}), X_{t+1/2} - x^* \rangle \\
\chi_{t+1/2} &\coloneqq 4\gamma_t^2 \|U_t\|_*^2 & \chi_{t+1} &\coloneqq 4\gamma_t^2 \|U_{t+1/2}\|_*^2 \\
\xi_{t+1/2} &\coloneqq 0 & \xi_{t+1} &\coloneqq -\gamma_t \langle U_{t+1/2}, X_{t+1/2} - x^* \rangle \\
D_t &\coloneqq D(x^*, X_t) + \phi_t & D_{t+1/2} &\coloneqq D_t - \zeta_t + \chi_{t+1/2} + \xi_{t+1/2}\,.
\end{aligned}$$

With these definitions, for any $t \in \mathbb{N}^*,\ t \geq 1$,

$$\begin{aligned}
D_{t+1/2} &\leq D_t - \zeta_t + \chi_{t+1/2} + \xi_{t+1/2} \\
D_{t+1} &\leq D_{t+1/2} - \zeta_{t+1/2} + \chi_{t+1} + \xi_{t+1}
\end{aligned}$$

where the first inequality comes directly from the definition while the second one comes from the descent inequality of Lemma B.2.

Now define the events, for $t \in \frac{1}{2}\mathbb{N}^*,\ t \geq 1$,

$$I_t = \{D_1 \leq C/2\} \cap \bigcap_{s \in \frac{1}{2}\mathbb{N} : 3/2 \leq s \leq t} \{D_s \leq C\} \cap \{\chi_s^2 \leq C/4\}\,.$$

We first show, by induction on $t \in \frac{1}{2}\mathbb{N}^*,\ t \geq 1$, that, $I_t \subset \{X_s \in \mathcal{U}, s = \frac{1}{2}, 1, \ldots, t - \frac{1}{2}, t\}$.

Initialization: For $t = 1$, the fact that $I_1 \subset \{X_{1/2} \in \mathcal{U}\}$ comes from the conditioning that $D(x^*, X_1) + \phi_1 \leq C/2$ and by definition of ϕ_1 and the strong convexity of h (Lemma A.1),

$$\|X_{1/2} - x^*\|^2 \leq 2(\|X_1 - x^*\|^2 + \|X_1 - X_{1/2}\|^2) \leq 4(D(x^*, X_1) + \phi_1) \leq 2C \leq r^2.$$

To show that $X_1 \in \mathcal{U}$, use the strong convexity of h (Lemma A.1) to get that, on I_1,

$$\|X_1 - x^*\|^2 \leq 2D(x^*, X_1) \leq 2D_1 \leq C \leq r^2\,.$$

Induction step for iterates: Assume that, for some $t \geq 2$, $I_{t-1/2} \subset \{X_s \in \mathcal{U}, s = \frac{1}{2}, 1, \ldots, t - 1, t - \frac{1}{2}\}$. By definition, the sequence of events (I_s) is non-increasing so that, $I_t \subset \{X_s \in \mathcal{U}, s = \frac{1}{2}, 1, \ldots, t - 1, t - \frac{1}{2}\}$. Hence, we only have to show that $I_t \subset \{X_t \in \mathcal{U}\}$. But, by definition, $I_t \subset \{D_t \leq C\}$. Again, using the strong convexity of h (Lemma A.1), on I_t,

$$\|X_t - x^*\|^2 \leq 2D(x^*, X_t) \leq 2D_t \leq 2C \leq r^2\,.$$

Induction step for half-iterates: Assume that, for some $t \geq 1$, $I_t \subset \{X_s \in \mathcal{U}, s = \frac{1}{2}, 1, \dots, t-1, t\}$. By definition, the sequence of events (I_s) is non-increasing so that, $I_{t+1/2} \subset \{X_s \in \mathcal{U}, s = \frac{1}{2}, 1, \dots, t-1, t\}$. So we focus on showing that $I_{t+1/2} \subset \{X_{t+1/2} \in \mathcal{U}\}$. For this, apply the first statement of Lemma A.3 with $(x, p, y, y') \leftarrow (X_t, x^*, -\gamma_t(v(X_{t-1/2}) + U_t), -\gamma_t v(x^*))$,

$$D(x^*, X_{t+1/2}) \leq D(x^*, X_t) - \gamma_t \langle v(X_{t-1/2}) + U_t - v(x^*), X_{t+1/2} - x^* \rangle \,,$$

and apply Young's inequality twice to get

$$\begin{aligned} D(x^*, X_{t+1/2}) &\leq D(x^*, X_t) + \gamma_t^2 \|v(X_{t-1/2}) + U_t - v(x^*)\|_*^2 + \frac{1}{4}\|X_{t+1/2} - x^*\|^2 \\ &\leq D(x^*, X_t) + 2\gamma_t^2 \|v(X_{t-1/2}) - v(x^*)\|_*^2 + 2\gamma_t^2 \|U_t\|^2 + \frac{1}{4}\|X_{t+1/2} - x^*\|^2 \,. \end{aligned}$$

By the strong convexity of h (Lemma A.1) and the Lipschitz continuity of v (Assumption 1),

$$\frac{1}{4}\|X_{t+1/2} - x^*\|^2 \leq D(x^*, X_t) + 2\gamma_t^2 L^2 \|X_{t-1/2} - x^*\|^2 + 2\gamma_t^2 \|U_t\|^2 \,.$$

But, by definition, on $I_{t+1/2}$, $D(x^*, X_t) \leq C$, $X_{t-1/2}$ is in $\mathcal{U}$ and $\chi_{t+1/2} = 4\gamma_t^2 \|U_t\|_*^2 \leq C/4$. Therefore,

$$\frac{1}{4}\|X_{t+1/2} - x^*\|^2 \leq \left(1 + \frac{1}{8}\right) C + 2\gamma_t^2 L^2 r^2 \,.$$

Using the definition of C and the bound on the step-size $\gamma_t \leq 1/(4L)$,

$$\|X_{t+1/2} - x^*\|^2 \leq \frac{r^2}{2} + \frac{r^2}{2} = r^2 \,,$$

which concludes the induction step.

We now verify the assumptions needed to apply Lemma B.3:

(i) For $t \in \frac{1}{2}\mathbb{N}^*$, $t \geq 1$, $\zeta_t \mathbb{1}_{I_t} \geq 0$. If $t \in \mathbb{N}^*$, this is trivial as $\zeta_t = 0$. Now, fix $t \in \mathbb{N}^*$, $\zeta_{t+1/2} = \gamma_t \langle v(X_{t+1/2}), X_{t+1/2} - x^* \rangle$. But, on $I_{t+1/2}$, $X_{t+1/2} \in \mathcal{U}$ and so, by monotonicity of v (Assumption 2), $\langle v(X_{t+1/2}), X_{t+1/2} - x^* \rangle \geq 0$ on $I_{t+1/2}$.

(ii) For $t \in \frac{1}{2}\mathbb{N}^*$, $t \geq 1$, $\mathbb{E}\left[\xi_{t+1/2} \,\middle|\, \mathcal{F}_t\right] \mathbb{1}_{I_t} = 0$. There is only something to prove when t is of the form $t + \frac{1}{2}$ with $t \in \mathbb{N}^*$. In this case, as $X_{t+1/2}$ is $\mathcal{F}_{t+1/2}$ measurable,

$$\mathbb{E}\left[\xi_{t+1} \mid \mathcal{F}_{t+1/2}\right] = -\gamma_t \langle \mathbb{E}\left[U_{t+1/2} \,\middle|\, \mathcal{F}_{t+1/2}\right], X_{t+1/2} - x^* \rangle = 0 \,.$$

(iii) $\sum_{t \in \frac{1}{2}\mathbb{N},\, t \geq 1} \mathbb{E}\left(\left(\xi_{t+1/2}^2 + \chi_{t+1/2}\right) \mathbb{1}_{I_t}\right) \leq \delta\epsilon\, \mathbb{P}(I_1)$ with $\epsilon = \min(C^2/16, C/4) = \min(C/4, 1)C/4$ and $\delta = \frac{\sigma^2(8+r^2)\Gamma}{\epsilon}$. For this let us first bound each of the terms involved individually. For $t \in \mathbb{N}^*$, by assumption on the noise and as I_t is $\mathcal{F}_t$ measurable, and so $\mathcal{F}_{t+1/2}$ measurable,

$$\begin{aligned} \mathbb{E}(\chi_{t+1/2}\mathbb{1}_{I_t}) &= 4\gamma_t^2\, \mathbb{E}\left(\|U_{t+1/2}\|_*^2 \mathbb{1}_{I_t}\right) \\ &= 4\gamma_t^2\, \mathbb{E}\left(\mathbb{E}\left[\|U_{t+1/2}\|_*^2 \,\middle|\, \mathcal{F}_{t+1/2}\right] \mathbb{1}_{I_t}\right) \\ &\leq 4\gamma_t^2 \sigma^2\, \mathbb{P}(I_t) \\ &\leq 4\gamma_t^2 \sigma^2\, \mathbb{P}(I_1) \,, \end{aligned}$$

where we used that the sequence of events $(I_s)_s$ is non-increasing. Likewise, $\mathbb{E}(\chi_{t+1}\mathbb{1}_{I_{t+1/2}}) \leq 4\gamma_t^2\sigma^2\,\mathbb{P}(I_1)$. Now, by definition of the dual norm,

$$\begin{aligned}\mathbb{E}(\xi_{t+1}^2\mathbb{1}_{I_{t+1/2}}) &= \gamma_t^2\,\mathbb{E}\left(\langle U_{t+1/2}, X_{t+1/2}-x^*\rangle^2\mathbb{1}_{I_{t+1/2}}\right)\\ &\leq \gamma_t^2\,\mathbb{E}\left(\|U_{t+1/2}\|_*^2\|X_{t+1/2}-x^*\|^2\mathbb{1}_{I_{t+1/2}}\right)\\ &\leq \gamma_t^2 r^2\,\mathbb{E}\left(\|U_{t+1/2}\|_*^2\mathbb{1}_{I_{t+1/2}}\right),\end{aligned}$$

where we used that $I_{t+1/2} \subset \{X_{t+1/2} \in \mathcal{U}\}$. Next, by the law of total expectation and since $I_{t+1/2}$ is $\mathcal{F}_{t+1/2}$ measurable,

$$\begin{aligned}\mathbb{E}(\xi_{t+1}^2\mathbb{1}_{I_{t+1/2}}) &\leq \gamma_t^2 r^2\,\mathbb{E}\left(\mathbb{E}\left[\,\|U_{t+1/2}\|_*^2 \,\middle|\, \mathcal{F}_{t+1/2}\right]\mathbb{1}_{I_{t+1/2}}\right)\\ &\leq \gamma_t^2 r^2\sigma^2\,\mathbb{P}(I_{t+1/2})\\ &\leq \gamma_t^2 r^2\sigma^2\,\mathbb{P}(I_1)\,.\end{aligned}$$

Combining these bounds, we get,

$$\begin{aligned}&\sum_{t\in\frac{1}{2}\mathbb{N},\ t\geq 1}\mathbb{E}\left(\left(\xi_{t+1/2}^2+\chi_{t+1/2}\right)\mathbb{1}_{I_t}\right)\\ &= \sum_{t\in\mathbb{N},\ t\geq 1}\mathbb{E}\left(\xi_{t+1}^2\mathbb{1}_{I_{t+1/2}}\right)+\sum_{t\in\mathbb{N},\ t\geq 1}\mathbb{E}\left(\chi_{t+1/2}^2\mathbb{1}_{I_t}\right)+\sum_{t\in\mathbb{N},\ t\geq 1}\mathbb{E}\left(\chi_{t+1}^2\mathbb{1}_{I_{t+1/2}}\right)\\ &\leq \sigma^2(8+r^2)\,\mathbb{P}(I_1)\sum_{t=1}^{+\infty}\gamma_t^2\\ &\leq \underbrace{\frac{\sigma^2(8+r^2)\Gamma}{\epsilon}}_{=\delta}\epsilon\,\mathbb{P}(I_1)\,,\end{aligned}$$

which corresponds to the statement.

Hence, Hsieh et al. (2020, Lem. F.1) gives that,

$$\mathbb{P}\left[\bigcap_{t\in\frac{1}{2}\mathbb{N},\ t\geq 1} I_t \,\middle|\, I_1\right] \geq 1-\delta\,,$$

which implies our statement since $\bigcap_{t\in\frac{1}{2}\mathbb{N},\ t\geq 1} I_t \subset \{\forall t \geq \frac{1}{2},\ X_t \in \mathcal{U}\}$. ■

Appendix C. Convergence of optimistic mirror descent

Proof of Proposition 1.

1. Start from Lemma B.2 which gives us the descent inequality

$$\begin{aligned}D(x^*, X_{t+1}) + \phi_{t+1} \leq& D(x^*, X_t) + (1-\gamma_t\mu)\phi_t - \gamma_t\langle V_{t+1/2}, X_{t+1/2}-x^*\rangle \\ &+\left(4\gamma_t^2L^2-\frac{1}{2}\right)\|X_{t+1/2}-X_t\|^2 + 4\gamma_t^2(\|U_{t+1/2}\|_*^2+\|U_{t-1/2}\|_*^2)\end{aligned} \tag{C.1}$$

where $\phi_t = \frac{\gamma_{t-1}^2}{2}\|V_{t-1/2} - V_{t-3/2}\|^2$ for all $t \geq 2$, and $\phi_1 = \frac{1}{2}\|X_1 - X_{1/2}\|^2 = 0$ by assumption.

2. The first part of the result comes the stochastic stability lemma Lemma B.4 (which relies on Hsieh et al. (2020)), where we use that the stepsize choice is square-summable with

$$\sum_{t=1}^{+\infty} \gamma_t^2 = \gamma \leq \frac{\gamma^2}{(2\eta - 1)t_0^{2\eta-1}}. \tag{C.2}$$

3. We now focus on proving the rates of convergence. Take $r > 0$ small enough so that $\mathcal{U} \coloneqq \mathbb{B}(x^*, r) \cap \mathcal{K}$ is included in both $\mathcal{V}$ and $\mathcal{B}$.

   Define the event $\mathcal{E}_t = \{\forall 1 \leq s \leq t,\ X_{s+1/2} \in \mathcal{U}\}$ for $t \geq 0$. Note that, except for $t = 0$, $\mathcal{E}_t$ is $\mathcal{F}_{t+1/2}$ is measurable. On this event, we can apply Lemma A.6 with $x \leftarrow X_t$ to get,

$$\langle v(X_{t+1/2}), X_{t+1/2} - x^*\rangle \geq \frac{\mu}{2}\|X_t - x^*\|^2 - \mu\|X_{t+1/2} - X_t\|^2 .$$

   and using the Legendre exponent of h, we get

$$\langle v(X_{t+1/2}), X_{t+1/2} - x^*\rangle \geq \frac{\mu}{\kappa}D(x^*, X_t)^{1+\alpha} - \mu\|X_{t+1/2} - X_t\|^2 .$$

4. Combining this with the descent inequality above Eq. (C.1), and the fact that $\kappa \geq 1$,

$$\begin{aligned}
&(D(x^*, X_{t+1}) + \phi_{t+1})\mathbb{1}_{\mathcal{E}_t} \\
&\leq (D(x^*, X_t) - \frac{\mu\gamma_t}{\kappa}D(x^*, X_t)^{1+\alpha} + (1 - \frac{\gamma_t\mu}{\kappa})\phi_t)\mathbb{1}_{\mathcal{E}_t} - \gamma_t\langle U_{t+1/2}, X_{t+1/2} - x^*\rangle\mathbb{1}_{\mathcal{E}_t} \\
&\quad + \left(4\gamma_t^2L^2 + \mu\gamma_t - \frac{1}{2}\right)\|X_{t+1/2} - X_t\|^2\mathbb{1}_{\mathcal{E}_t} + 4\gamma_t^2(\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2)\mathbb{1}_{\mathcal{E}_t} .
\end{aligned}$$

   Now, the sequence of events $(\mathcal{E}_s)_{s\geq 0}$ is non-increasing, so $\mathbb{1}_{\mathcal{E}_t} \leq \mathbb{1}_{\mathcal{E}_{t-1}} \leq 1$. As a consequence,

$$\begin{aligned}
&\left(D(x^*, X_t) - \frac{\mu\gamma_t}{\kappa}D(x^*, X_t)^{1+\alpha} + (1 - \frac{\gamma_t\mu}{\kappa})\phi_t\right)\mathbb{1}_{\mathcal{E}_t} \\
&\leq \left(D(x^*, X_t) - \frac{\mu\gamma_t}{\kappa}D(x^*, X_t)^{1+\alpha} + (1 - \gamma_t\mu)\phi_t\right)\mathbb{1}_{\mathcal{E}_{t-1}} .
\end{aligned}$$

   Note that the term between parenthesis is always non-negative even when $\alpha > 0$.

   Moreover, by the choice of γ, $4\gamma_t^2L^2 + \mu\gamma_t - \frac{1}{2} \leq 0$. Therefore, the descent inequality can be simplified to give,

$$\begin{aligned}
&\left(D(x^*, X_{t+1}) + \phi_{t+1}\right)\mathbb{1}_{\mathcal{E}_t} \\
&\leq \left(D(x^*, X_t) - \frac{\mu\gamma_t}{\kappa}D(x^*, X_t)^{1+\alpha} + (1 - \frac{\gamma_t\mu}{\kappa})\phi_t\right)\mathbb{1}_{\mathcal{E}_{t-1}} - \gamma_t\langle U_{t+1/2}, X_{t+1/2} - x^*\rangle\mathbb{1}_{\mathcal{E}_t} \\
&\quad + 4\gamma_t^2(\|U_{t+1/2}\|_*^2 + \|U_{t-1/2}\|_*^2) .
\end{aligned}$$

   We now take the expectation of this inequality. For this, note that, as $\mathcal{E}_t$ is $\mathcal{F}_{t+1/2}$ measurable,

$$\begin{aligned}
\mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2} - x^*\rangle\mathbb{1}_{\mathcal{E}_t}\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle U_{t+1/2}, X_{t+1/2} - x^*\rangle\mathbb{1}_{\mathcal{E}_t} \,\middle|\, \mathcal{F}_{t+1/2}\right]\right] \\
&= \mathbb{E}\left[\langle\mathbb{E}\left[U_{t+1/2} \,\middle|\, \mathcal{F}_{t+1/2}\right], X_{t+1/2} - x^*\rangle\mathbb{1}_{\mathcal{E}_t}\right] \\
&= 0 .
\end{aligned}$$

As a consequence, and using the assumption on the noise,

$$\begin{aligned}&\mathbb{E}\left[(D(x^*, X_{t+1}) + \phi_{t+1})\mathbb{1}_{\mathcal{E}_t}\right]\\&\leq \mathbb{E}\left[\left(D(x^*, X_t) - \frac{\mu\gamma_t}{\kappa}D(x^*, X_t)^{1+\alpha} + \left(1 - \frac{\gamma_t\mu}{\kappa}\right)\phi_t\right)\mathbb{1}_{\mathcal{E}_{t-1}}\right] + 8\gamma_t^2\sigma^2\,.\end{aligned} \tag{C.3}$$

5. The final step will be to study the sequence (a_t) defined by $a_t = \mathbb{E}\left[(D(x^*, X_t) + \phi_t)\mathbb{1}_{\mathcal{E}_{t-1}}\right]$ for all $t \geq 1$. Indeed, a bound on $a_t = \mathbb{E}\left((D(x^*, X_t) + \phi_t)\mathbb{1}_{\mathcal{E}_{t-1}}\right)$ implies a bound on the quantity of interest,
$$\mathbb{E}\left(D(x^*, X_t) + \phi_t \mid \mathcal{E}_{\mathcal{U}}\right)\,.$$
As $\mathcal{E}_{\mathcal{U}} \subset \mathcal{E}_t$,
$$\begin{aligned}\mathbb{E}\left[D(x^*, X_t) + \phi_t \mid \mathcal{E}_{\mathcal{U}}\right] &= \frac{\mathbb{E}\left[(D(x^*, X_t) + \phi_t)\mathbb{1}_{\mathcal{E}}\right]}{\mathbb{P}[\mathcal{E}_{\mathcal{U}}]}\\&\leq \frac{\mathbb{E}\left[(D(x^*, X_t) + \phi_t)\mathbb{1}_{\mathcal{E}_t}\right]}{\mathbb{P}[\mathcal{E}_{\mathcal{U}}]}\end{aligned}$$
Now, the inequality Eq. (C.3) above can be rewritten as follows,
$$a_{t+1} \leq a_t - \frac{\mu\gamma_t}{\kappa}\,\mathbb{E}\left[\left(D(x^*, X_t)^{1+\alpha} + \phi_t\right)\mathbb{1}_{\mathcal{E}_{t-1}}\right] + 8\gamma_t^2\sigma^2\,. \tag{C.4}$$

6. Now, the behavior of this sequence depends heavily on the Legendre exponent α:
   - If $\alpha = 0$, (C.4) becomes
   $$a_{t+1} \leq \left(1 - \frac{\mu\gamma}{\kappa(t+t_0)^\eta}\right)a_t + \frac{8\gamma^2\sigma^2}{(t+t_0)^{2\eta}}\,.$$
   Then, for $\eta = 1$, Lemma A.8 can be applied with the additional assumption that $\gamma > \frac{\kappa}{\mu}$ to get that for any $T \geq 1$, $a_T = \mathcal{O}(1/T)$.
   For $\frac{1}{2} < \eta < 1$, Lemma A.9 can be directly apply to obtain $a_T = \mathcal{O}(1/T^\eta)$.
   - If $\alpha > 0$, a little more work has to be done before on (C.4) before concluding. First, (C.4) implies that for any $t \geq 1$, $a_t \leq a_0 + 8\sigma^2\sum_{t=1}^{\infty}\gamma_t^2$ and since we assume $D(x^*, X_1) + \phi_1 = D(x^*, X_1) \leq \frac{r^2}{12}$ we have that
   $$\begin{aligned}\mathbb{E}\left[\phi_t\mathbb{1}_{\mathcal{E}_{t-1}}\right] &\leq \mathbb{E}\left[(D(x^*, X_t) + \phi_t)\mathbb{1}_{\mathcal{E}_{t-1}}\right] = a_t\\&\leq \mathbb{E}\left[(D(x^*, X_1) + \phi_1)\mathbb{1}_{\mathcal{E}_0}\right] + 8\sigma^2\sum_{t=1}^{\infty}\gamma_t^2\\&\leq \underbrace{\frac{r^2}{12} + \frac{8\sigma^2\gamma^2}{(2\eta-1)t_0^{2\eta-1}}}_{:=\Phi}\,.\end{aligned}$$

As a consequence, $-\mathbb{E}\left[\phi_t \mathbb{1}_{\mathcal{E}_{t-1}}\right] \leq -\mathbb{E}\left[\phi_t \mathbb{1}_{\mathcal{E}_{t-1}}\right]^{1+\alpha} / \Phi^\alpha$, and thus (C.4) gives us

$$\begin{aligned} a_{t+1} &\leq a_t - \frac{\mu\gamma_t}{\kappa}\,\mathbb{E}\left[\left(D(x^*, X_t)^{1+\alpha} + \frac{1}{\Phi^\alpha}\phi_t^{1+\alpha}\right)\mathbb{1}_{\mathcal{E}_{t-1}}\right] + 8\gamma_t^2\sigma^2 \\ &\leq a_t - \frac{\mu\gamma_t}{2^\alpha \max(1,\Phi^\alpha)\kappa}\,\mathbb{E}\left[\left(D(x^*, X_t) + \phi_t\right)^{1+\alpha}\mathbb{1}_{\mathcal{E}_{t-1}}\right] + 8\gamma_t^2\sigma^2 \\ &\leq a_t - \frac{\mu\gamma_t}{2^\alpha \max(1,\Phi^\alpha)\kappa} a_t^{1+\alpha} + 8\gamma_t^2\sigma^2 \end{aligned}$$

where the second inequality comes from the fact that $(x+y)^{1+\alpha}/2^\alpha \leq x^{1+\alpha} + y^{1+\alpha}$ for positive x, y and the last one from Jensen inequality applied to the convex function $x \mapsto x^{1+\alpha}$.

We will now apply one of Lemmas A.11 and A.12 to the sequence (a_t). To make this step clear, let us introduce the same notations as these lemmas. Define

$$q \coloneqq \frac{\mu\gamma}{2^\alpha \max(1,\Phi^\alpha)\kappa} \quad \text{and} \quad q' \coloneqq 8\sigma^2\gamma^2\,.$$

With these notations, the descent inequality can be rewritten as,

$$a_{t+1} \leq a_t - \frac{q}{(t+t_0)^\eta} a_t^{1+\alpha} + \frac{q'}{(t+t_0)^{2\eta}}\,.$$

Both Lemmas A.11 and A.12 require that,

$$q' q^{1/\alpha} \leq c(\eta,\alpha) \iff \gamma^{2+\frac{1}{\alpha}} \leq \frac{c(1,\alpha)}{4\sigma^2(\mu/\kappa)^{1/\alpha}} \max(1,\Phi)\,.$$

Finally, we distinguish two cases.

– If $\eta \geq \frac{1+\alpha}{1+2\alpha}$, we apply Lemma A.11 to get that, for any $T \geq 1$,

$$a_T \leq \frac{a_1 + b}{\left(1 + \alpha(a_1+b)^\alpha 2^{-\alpha} \sum_{t=1}^{T-1} \frac{q}{(t+t_0)^\eta}\right)^{1/\alpha}}\,,$$

where $b = \left(\frac{1-2^{1-2\eta}}{(1+\alpha)q}\right)^{\frac{1}{\alpha}}$. Using asymptotic notations, this simply means that,

$$a_T = \mathcal{O}\left(\left(\sum_{t=1}^{T-1} \frac{1}{(t+t_0)^\eta}\right)^{-1/\alpha}\right).$$

The final bound on a_T now comes from the fact that,

$$\sum_{t=1}^{T-1} \frac{1}{(t+t_0)^\eta} = \begin{cases} \Theta\left(T^{1-\eta}\right) & \text{if } \eta < 1 \\ \Theta\left(\log T\right) & \text{if } \eta = 1\,. \end{cases}$$

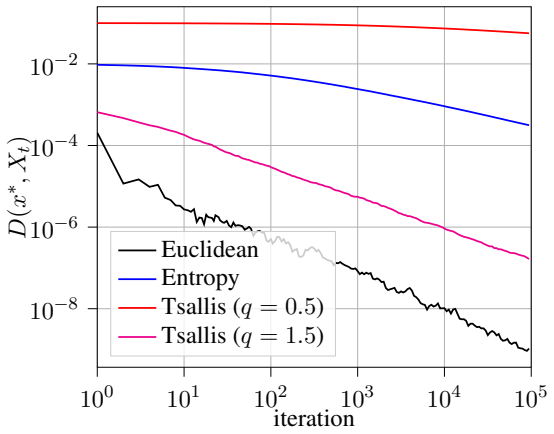


Figure 2: Convergence of OMD with different DGFs in terms of Bregman divergence to the solution.

– If $\eta \leq \frac{1+\alpha}{1+2\alpha}$, we apply Lemma A.12 to get that, for any $T \geq 1$,

$$a_T \leq \frac{a_1}{\left(1 + \alpha a_1^\alpha \sum_{t=1}^{T-1} \frac{q}{(t+t_0)^\eta}\right)^{1/\alpha}} + \frac{1}{((1+\alpha)q)^{\frac{1}{\alpha}}(t+t_0)^\eta}.$$

In other words, the sequence (a_t) satisfies, as $\eta < 1$,

$$a_T = \mathcal{O}\left(\frac{1}{T^{(1-\eta)/\alpha}}\right) + \mathcal{O}\left(\frac{1}{T^\eta}\right).$$

■

Appendix D. Illustration

We describe here the setup used in the illustration of Theorem 1 and provide additional plots.

We consider a simple $1D$ example with $\mathcal{K} = [0, +\infty)$ and $v(x) = x$; the solution of the associated variational inequality is thus $x^* = 0$. It is direct to see that v is $L = 1$ Lipschitz continuous and $\mu = 1$ strongly monotone. We consider the three Bregman regularizers of our running examples on $\mathcal{K} = [0, +\infty)$:

- *Euclidean projection* (Example 3.1): $h(x) = x^2/2$ for which $D(x^*, x) = x^2/2$, and $\beta = 0$;
- *Negative entropy* (Example 3.2): $h(x) = x \log x$ for which $D(x^*, x) = x$ and $\beta = 1/2$ at x^*;
- *Tsallis entropy* (Example 3.3): $h(x) = -\frac{1}{q(1-q)}x^q$ for which $D(x^*, x) = x^q/q$ and $\beta = \max(0, 1 - q/2)$ at x^*. To show different behaviors we consider $q = 0.5$ ($\beta = 0.75$) and $q = 1.5$ ($\beta = 0.25$).

For each of these regularizers, we run OMD initialized with $X_1 = X_{1/2} = 0.1$, (U_t) an i.i.d. Gaussian noise process with mean 0 and variance $\sigma^2 = 10^{-4}$. We choose our stepsize sequence as $\gamma_t = \frac{1}{t^\eta}$ with η as prescribed by Theorem 1. We display the results averaged over 100 runs.

We find out that the observed rates match the theory:

Regularizer	β	γ_t	Theoretical rate	Observed rate (by regression)
Euclidean	0	$\frac{1}{t}$	$\frac{1}{t}$	$\frac{1}{t^{0.99}}$
Entropy	0.5	$\frac{1}{t^{0.5+\epsilon}}$	$\frac{1}{t^{0.5+\epsilon}}$	$\frac{1}{t^{0.48}}$
Tsallis ($q = 0.5$)	0.75	$\frac{1}{t^{0.5+\epsilon}}$	$\frac{1}{t^{0.1666}}$	$\frac{1}{t^{0.13}}$
Tsallis ($q = 1.5$)	0.25	$\frac{1}{t^{0.75}}$	$\frac{1}{t^{0.75}}$	$\frac{1}{t^{0.71}}$