On the Convergence of Single-Call Stochastic Extra-Gradient Methods

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos

On the Convergence of Single-Call Stochastic Extra-Gradient Methods

Yu-Guan Hsieh
Univ. Grenoble Alpes, LJK and ENS Paris
38000 Grenoble, France.
yu-guan.hsieh@ens.fr

Franck Iutzeler
Univ. Grenoble Alpes, LJK
38000 Grenoble, France.
franck.iutzeler@univ-grenoble-alpes.fr

Jérôme Malick
CNRS, LJK
38000 Grenoble, France.
jerome.malick@univ-grenoble-alpes.fr

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

Abstract

Variational inequalities have recently attracted considerable interest in machine learning as a flexible paradigm for models that go beyond ordinary loss function minimization (such as generative adversarial networks and related deep learning systems). In this setting, the optimal $\mathcal{O}(1/t)$ convergence rate for solving smooth monotone variational inequalities is achieved by the Extra-Gradient (EG) algorithm and its variants. Aiming to alleviate the cost of an extra gradient step per iteration (which can become quite substantial in deep learning applications), several algorithms have been proposed as surrogates to Extra-Gradient with a *single* oracle call per iteration. In this paper, we develop a synthetic view of such algorithms, and we complement the existing literature by showing that they retain a $\mathcal{O}(1/t)$ ergodic convergence rate in smooth, deterministic problems. Subsequently, beyond the monotone deterministic case, we also show that the last iterate of single-call, *stochastic* extra-gradient methods still enjoys a $\mathcal{O}(1/t)$ local convergence rate to solutions of *non-monotone* variational inequalities that satisfy a second-order sufficient condition.

1 Introduction

Deep learning is arguably the fastest-growing field in artificial intelligence: its applications range from image recognition and natural language processing to medical anomaly detection, drug discovery, and most fields where computers are required to make sense of massive amounts of data. In turn, this has spearheaded a prolific research thrust in optimization theory with the twofold aim of demystifying the successes of deep learning models and of providing novel methods to overcome their failures.

Introduced by Goodfellow et al. [21], generative adversarial networks (GANs) have become the youngest torchbearers of the deep learning revolution and have occupied the forefront of this drive in more ways than one. First, the adversarial training of deep neural nets has given rise to new challenges regarding the efficient allocation of parallelizable resources, the compatibility of the

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

	Lipschitz		Lipschitz + Strong	
	Ergodic	Last Iterate	Ergodic	Last Iterate
Deterministic	$\boxed{1/t}$	Unknown	$1/t$	$e^{-\rho t}$ [19, 26, 32]
Stochastic	$1/\sqrt{t}$ [14, 19]	Unknown	$\boxed{1/t}$	$\boxed{1/t}$

Table 1: The best known global convergence rates for single-call extra-gradient methods in monotone VI problems; logarithmic factors ignored throughout. A box indicates a contribution from this paper.

chosen architectures, etc. Second, the loss landscape in GANs is no longer that of a minimization problem but that of a zero-sum, min-max game – or, more generally, a *variational inequality* (VI).

Variational inequalities are a flexible and widely studied framework in optimization which, among others, incorporates minimization, saddle-point, Nash equilibrium, and fixed point problems. As such, there is an extensive literature devoted to solving variational inequalities in different contexts; for an introduction, see [4, 18] and references therein. In particular, in the setting of monotone variational inequalities with Lipschitz continuous operators, it is well known that the optimal rate of convergence is $\mathcal{O}(1/t)$, and that this rate is achieved by the Extra-Gradient (EG) algorithm of Korpelevich [24] and its Bregman variant, the Mirror-Prox (MP) algorithm of Nemirovski [33].[1]

These algorithms require two projections and two oracle calls per iteration, so they are more costly than standard Forward-Backward / descent methods. As a result, there are two complementary strands of literature aiming to reduce one (or both) of these cost multipliers – that is, the number of projections and/or the number of oracle calls per iteration. The first class contains algorithms like the Forward-Backward-Forward (FBF) method of Tseng [44], while the second focuses on gradient extrapolation mechanisms like Popov's modified Arrow–Hurwicz algorithm [38].

In deep learning, the latter direction has attracted considerably more interest than the former. The main reason for this is that neural net training often does not involve constraints (and, when it does, they are relatively cheap to handle). On the other hand, gradient calculations can become very costly, so a decrease in the number of oracle calls could offer significant practical benefits. In view of this, our aim in this paper is (*i*) to develop a synthetic approach to methods that retain the anticipatory properties of the Extra-Gradient algorithm while making a single oracle call per iteration; and (*ii*) to derive quantitative convergence results for such *single-call extra-gradient* (1-EG) algorithms.

Our contributions. Our first contribution complements the existing literature (reviewed below and in Section 3) by showing that the class of 1-EG algorithms under study attains the optimal $\mathcal{O}(1/t)$ convergence rate of the two-call method in deterministic variational inequalities with a monotone, Lipschitz continuous operator. Subsequently, we show that this rate is also achieved in *stochastic* variational inequalities with strongly monotone operators provided that the optimizer has access to an oracle with bounded variance (but not necessarily bounded second moments).

Importantly, this stochastic result concerns both the method's "ergodic average" (a weighted average of the sequence of points generated by the algorithm) as well as its "last iterate" (the last generated point). The reason for this dual focus is that averaging can be very useful in convex/monotone landscapes, but it is not as beneficial in non-monotone problems (where Jensen's inequality does not apply). On that account, last-iterate convergence results comprise an essential stepping stone for venturing beyond monotone problems.

Armed with these encouraging results, we then focus on *non-monotone* problems and show that, with high probability, the method's last iterate exhibits a $\mathcal{O}(1/t)$ local convergence rate to solutions of non-monotone variational inequalities that satisfy a second-order sufficient condition. To the best of our knowledge, this is the first convergence rate guarantee of this type for stochastic, non-monotone variational inequalities.

Related work. The prominence of Extra-Gradient/Mirror-Prox methods in solving variational inequalities and saddle-point problems has given rise to a vast corpus of literature which we cannot hope to do justice here. Especially in the context of adversarial networks, there has been a flurry

[1] Korpelevich [24] proved the method's asymptotic convergence for pseudomonotone variational inequalities. The $\mathcal{O}(1/t)$ convergence rate was later established by Nemirovski [33] with ergodic averaging.

of recent activity relating variants of the Extra-Gradient algorithm to GAN training, see e.g., [9, 15, 19, 20, 25, 29, 45] and references therein. For concreteness, we focus here on algorithms with a single-call structure and refer the reader to Sections 3–5 for additional details.

The first variant of Extra-Gradient with a single oracle call per iteration dates back to Popov [38]. This algorithm was subsequently studied by, among others, Chiang et al. [10], Rakhlin and Sridharan [39, 40] and Gidel et al. [19]; see also [14, 26] for a "reflected" variant, [15, 31, 32, 37] for an "optimistic" one, and Section 3 for a discussion of the differences between these variants. In the context of deterministic, strongly monotone variational inequalities with Lipschitz continuous operators, the last iterate of the method was shown to exhibit a geometric convergence rate [19, 26, 32, 43]; similar geometric convergence results also extend to bilinear saddle-point problems [19, 37, 43], even though the operator involved is not strongly monotone. In turn, this implies the convergence of the method's ergodic average, but at a $\mathcal{O}(1/t)$ rate (because of the hysteresis of the average). In view of this, the fact that 1-EG methods retain the optimal $\mathcal{O}(1/t)$ convergence rate in deterministic variational inequalities without strong monotonicity assumptions closes an important gap in the literature.[2]

At the local level, the geometric convergence results discussed above echo a surge of interest in local convergence guarantees of optimization algorithms applied to games and saddle-point problems, see e.g., [1, 3, 16, 25] and references therein. In more detail, Liang and Stokes [25] proved local geometric convergence for several algorithms in possibly non-monotone saddle-point problems under a local smoothness condition. In a similar vein, Daskalakis and Panageas [16] analyzed the limit points of (optimistic) gradient descent, and showed that local saddle points are stable stationary points; subsequently, Adolphs et al. [1] and Mazumdar et al. [28] proposed a class of algorithms that eliminate stationary points which are not local Nash equilibria.

Geometric convergence results of this type are inherently deterministic because they rely on an associated resolvent operator being firmly nonexpansive – or, equivalently, rely on the use of the center manifold theorem. In a stochastic setting, these techniques are no longer applicable because the contraction property cannot be maintained in the presence of noise; in fact, unless the problem at hand is amenable to variance reduction – e.g., as in [6, 9, 22] – geometric convergence is not possible if the noise process is even weakly isotropic. Instead, for monotone problems, Cui and Shanbhag [14] and Gidel et al. [19] showed that the ergodic average of the method attains a $\mathcal{O}(1/\sqrt{t})$ convergence rate. Our global convergence results for stochastic variational inequalities improve this rate to $\mathcal{O}(1/t)$ in strongly monotone variational inequalities for both the method's ergodic average and its last iterate. In the same light, our local $\mathcal{O}(1/t)$ convergence results for *non-monotone* variational inequalities provide a key extension of local, deterministic convergence results to a fully stochastic setting, all the while retaining the fastest convergence rate for monotone variational inequalities.

For convenience, our contributions relative to the state of the art are summarized in Table 1.

2 Problem setup and blanket assumptions

Variational inequalities. We begin by presenting the basic variational inequality framework that we will consider throughout the sequel. To that end, let $\mathcal{X}$ be a nonempty closed convex subset of $\mathbb{R}^d$, and let $V\colon \mathbb{R}^d \to \mathbb{R}^d$ be a single-valued operator on $\mathbb{R}^d$. In its most general form, the *variational inequality* (VI) problem associated to V and $\mathcal{X}$ can be stated as:

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

To provide some intuition about (VI), we discuss two important examples below:

Example 1 (Loss minimization)**.** Suppose that $V = \nabla f$ for some smooth loss function f on $\mathcal{X} = \mathbb{R}^d$. Then, $x^\star \in \mathcal{X}$ is a solution to (VI) if and only if $\nabla f(x^\star) = 0$, i.e., if and only if $x^\star$ is a critical point of f. Of course, if f is convex, any such solution is a global minimizer. □

Example 2 (Min-max optimization)**.** Suppose that $\mathcal{X}$ decomposes as $\mathcal{X} = \Theta \times \Phi$ with $\Theta = \mathbb{R}^{d_1}$, $\Phi = \mathbb{R}^{d_2}$, and assume $V = (\nabla_\theta \mathcal{L}, -\nabla_\phi \mathcal{L})$ for some smooth function $\mathcal{L}(\theta, \phi)$, $\theta \in \Theta$, $\phi \in \Phi$. As in

[2] A few weeks after the submission of our paper, we were made aware of a very recent preprint by Mokhtari et al. [31] which also establishes a $\mathcal{O}(1/t)$ convergence rate for the algorithm's "optimistic" variant in saddle-point problems (in terms of the Nikaido–Isoda gap function). To the best of our knowledge, this is the closest result to our own in the literature.

Example 1 above, the solutions to (VI) correspond to the critical points of $\mathcal{L}$; if, in addition, $\mathcal{L}$ is convex-concave, any solution $x^\star = (\theta^\star, \phi^\star)$ of (VI) is a global *saddle-point*, i.e.,

$$\mathcal{L}(\theta^\star, \phi) \leq \mathcal{L}(\theta^\star, \phi^\star) \leq \mathcal{L}(\theta, \phi^\star) \quad \text{for all } \theta \in \Theta \text{ and all } \phi \in \Phi.$$

Given the original formulation of GANs as (stochastic) saddle-point problems [21], this observation has been at the core of a vigorous literature at the interface between optimization, game theory, and deep learning, see e.g., [9, 15, 19, 25, 29, 37, 45] and references therein. □

The operator analogue of convexity for a function is *monotonicity*, i.e.,

$$\langle V(x') - V(x), x' - x \rangle \geq 0 \quad \text{for all } x, x' \in \mathbb{R}^d.$$

Specifically, when $V = \nabla f$ for some sufficiently smooth function f, this condition is equivalent to f being convex [4]. In this case, following Nesterov [35, 36] and Juditsky et al. [23], the quality of a candidate solution $\hat{x} \in \mathcal{X}$ can be assessed via the so-called *error* (or *merit*) *function*

$$\mathrm{Err}(\hat{x}) = \sup_{x \in \mathcal{X}} \langle V(x), \hat{x} - x \rangle$$

and/or its restricted variant

$$\mathrm{Err}_R(\hat{x}) = \max_{x \in \mathcal{X}_R} \langle V(x), \hat{x} - x \rangle,$$

where $\mathcal{X}_R \equiv \mathcal{X} \cap \mathbb{B}_R(0) = \{x \in \mathcal{X} : \|x\| \leq R\}$ denotes the "restricted domain" of the problem. More precisely, we have the following basic result.

Lemma 1 (Nesterov, 2007)**.** *Assume V is monotone. If $x^\star$ is a solution of (VI), we have $\mathrm{Err}(x^\star) = 0$ and $\mathrm{Err}_R(x^\star) = 0$ for all sufficiently large R. Conversely, if $\mathrm{Err}_R(\hat{x}) = 0$ for large enough $R > 0$ and some $\hat{x} \in \mathcal{X}_R$, then $\hat{x}$ is a solution of (VI).*

In light of this result, Err and Err_R will be among our principal measures of convergence in the sequel.

Blanket assumptions. With all this in hand, we present below the main assumptions that will underlie the bulk of the analysis to follow.

Assumption 1. The solution set $\mathcal{X}^\star$ of (VI) is nonempty.

Assumption 2. The operator V is β-Lipschitz continuous, i.e.,

$$\|V(x') - V(x)\| \leq \beta \|x' - x\| \quad \text{for all } x, x' \in \mathbb{R}^d.$$

Assumption 3. The operator V is monotone.

In some cases, we will also strengthen Assumption 3 to:

Assumption 3(s). The operator V is α-strongly monotone, i.e.,

$$\langle V(x') - V(x), x' - x \rangle \geq \alpha \|x' - x\|^2 \quad \text{for some } \alpha > 0 \text{ and all } x, x' \in \mathbb{R}^d.$$

Throughout our paper, we will be interested in sequences of points $X_t \in \mathcal{X}$ generated by algorithms that can access the operator V via a *stochastic oracle* [34].[3] Formally, this is a black-box mechanism which, when called at $X_t \in \mathcal{X}$, returns the estimate

$$V_t = V(X_t) + Z_t, \tag{1}$$

where $Z_t \in \mathbb{R}^d$ is an additive noise variable satisfying the following hypotheses:

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

$$b) \ \textit{Finite variance:} \quad \mathbb{E}[\|Z_t\|^2 \mid \mathcal{F}_t] \leq \sigma^2. \tag{2b}$$

In the above, $\mathcal{F}_t$ denotes the history (natural filtration) of X_t, so X_t is adapted to $\mathcal{F}_t$ by definition; on the other hand, since the t-th instance of Z_t is generated randomly from X_t, Z_t is *not* adapted to $\mathcal{F}_t$. Obviously, if $\sigma^2 = 0$, we have the deterministic, *perfect feedback* case $V_t = V(X_t)$.

[3] Depending on the algorithm, the sequence index t may take positive integer or half-integer values (or both).

3 Algorithms

The Extra-Gradient algorithm. In the general framework outlined in the previous section, the Extra-Gradient (EG) algorithm of Korpelevich [24] can be stated in recursive form as

$$\begin{aligned} X_{t+1/2} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_t) \\ X_{t+1} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_{t+1/2}) \end{aligned} \tag{EG}$$

where $\Pi_{\mathcal{X}}(y) \coloneqq \arg\min_{x\in\mathcal{X}} \|y - x\|$ denotes the Euclidean projection of $y \in \mathbb{R}^d$ onto the closed convex set $\mathcal{X}$ and $\gamma_t > 0$ is a variable step-size sequence. Using this formulation as a starting point, the main idea behind the method can be described as follows: at each $t = 1, 2, \dots$, the oracle is called at the algorithm's current – or *base* – state X_t to generate an intermediate – or *leading* – state $X_{t+1/2}$; subsequently, the base state X_t is updated to X_{t+1} using gradient information from the leading state $X_{t+1/2}$, and the process repeats. Heuristically, the extra oracle call allows the algorithm to "anticipate" the landscape of V and, in so doing, to achieve improved convergence results relative to standard projected gradient / forward-backward methods; for a detailed discussion, we refer the reader to [7, 18] and references therein.

Single-call variants of the Extra-Gradient algorithm. Given the significant computational overhead of gradient calculations, a key desideratum is to drop the second oracle call in (EG) while retaining the algorithm's "anticipatory" properties. In light of this, we will focus on methods that perform a *single* oracle call at the leading state $X_{t+1/2}$, but replace the update rule for $X_{t+1/2}$ (and, possibly, X_t as well) with a proxy that compensates for the missing gradient. Concretely, we will examine the following family of *single-call extra-gradient* (1-EG) algorithms:

1. *Past Extra-Gradient* (PEG) [10, 19, 38]:

$$\begin{aligned} X_{t+1/2} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_{t-1/2}) \\ X_{t+1} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_{t+1/2}) \end{aligned} \tag{PEG}$$

   [Proxy: use $V_{t-1/2}$ instead of V_t in the calculation of $X_{t+1/2}$]

2. *Reflected Gradient* (RG) [8, 14, 26]:

$$\begin{aligned} X_{t+1/2} &= X_t - (X_{t-1} - X_t) \\ X_{t+1} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_{t+1/2}) \end{aligned} \tag{RG}$$

   [Proxy: use $(X_{t-1} - X_t)/\gamma_t$ instead of V_t in the calculation of $X_{t+1/2}$; no projection]

3. *Optimistic Gradient* (OG) [15, 31, 32, 37]:

$$\begin{aligned} X_{t+1/2} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_{t-1/2}) \\ X_{t+1} &= X_{t+1/2} + \gamma_t V_{t-1/2} - \gamma_t V_{t+1/2} \end{aligned} \tag{OG}$$

   [Proxy: use $V_{t-1/2}$ instead of V_t in the calculation of $X_{t+1/2}$; use $X_{t+1/2} + \gamma_t V_{t-1/2}$ instead of X_t in the calculation of X_{t+1}; no projection]

These are the main algorithmic schemes that we will consider, so a few remarks are in order. First, given the extensive literature on the subject, this list is not exhaustive; see e.g., [31, 32, 37] for a generalization of (OG), [**?**] for a variant that employs averaging to update the algorithm's base state X_t, and [20] for a proxy defined via "negative momentum". Nevertheless, the algorithms presented above appear to be the most widely used single-call variants of (EG), and they illustrate very clearly the two principal mechanisms for approximating missing gradients: (*i*) using past gradients (as in the PEG and OG variants); and/or (*ii*) using a difference of successive states (as in the RG variant).

We also take this opportunity to provide some background and clear up some issues on terminology regarding the methods presented above. First, the idea of using past gradients dates back at least to Popov [38], who introduced (PEG) as a "modified Arrow–Hurwicz" method a few years after the original paper of Korpelevich [24]; the same algorithm is called "meta" in [10] and "extrapolation from the past" in [19] (but see also the note regarding optimism below). The terminology "Reflected

Gradient" and the precise formulation that we use here for (RG) is due to Malitsky [26]. The well-known primal-dual algorithm of Chambolle and Pock [8] can be seen as a one-sided, alternating variant of the method for saddle-point problems; see also [45] for a more recent take.

Finally, the terminology "optimistic" is due to Rakhlin and Sridharan [39, 40], who provided a unified view of (PEG) and (EG) based on the sequence of oracle vectors used to update the algorithm's leading state $X_{t+1/2}$.[4] Because the framework of [39, 40] encompasses two different algorithms, there is some danger of confusion regarding the use of the term "optimism"; in particular, both (EG) and (PEG) can be seen as instances of optimism. The specific formulation of (OG) that we present here is the projected version of the algorithm considered by Daskalakis et al. [15];[5] by contrast, the "optimistic" method of Mertikopoulos et al. [29] is equivalent to (EG) – not (PEG) or (OG).

The above shows that there can be a broad array of single-call extra-gradients methods depending on the specific proxy used to estimate the missing gradient, whether it is applied to the algorithm's base or leading state, when (or where) a projection operator is applied, etc. The contact point of all these algorithms is the unconstrained setting ($\mathcal{X} = \mathbb{R}^d$) where they are exactly equivalent:

Proposition 1. *Suppose that the* 1*-EG methods presented above share the same initialization,* $X_0 = X_1 \in \mathcal{X}$*,* $V_{1/2} = 0$*, and are run with the same, constant step-size* $\gamma_t \equiv \gamma$ *for all* $t \geq 1$*. If* $\mathcal{X} = \mathbb{R}^d$*, the generated iterates* X_t *coincide for all* $t \geq 1$*.*

The proof of this proposition follows by a simple rearrangement of the update rules for (PEG), (RG) and (OG), so we omit it. In the projected case, the 1-EG updates presented above are no longer equivalent – though, of course, they remain closely related.

4 Deterministic analysis

We begin with the deterministic analysis, i.e., when the optimizer receives oracle feedback of the form (1) with $\sigma = 0$. In terms of presentation, we keep the global and local cases separated and we interleave our results for the generated sequence X_t and its *ergodic average*. To streamline our presentation, we defer the details of the proofs to the paper's supplement and only discuss here the main ideas.

4.1 Global convergence

Our first result below shows that the algorithms under study achieve the optimal $\mathcal{O}(1/t)$ ergodic convergence rate in monotone problems with Lipschitz continuous operators.

Theorem 1. *Suppose that* V *satisfies Assumptions 1–3. Assume further that a* 1*-EG algorithm is run with perfect oracle feedback and a constant step-size* $\gamma < 1/(c\beta)$*, where* $c = 1 + \sqrt{2}$ *for the RG variant and* $c = 2$ *for the PEG and OG variants. Then, for all* $R > 0$*, we have*

$$\operatorname{Err}_R\left(\bar{X}_t\right) \leq \frac{R^2 + \|X_1 - X_{1/2}\|^2}{2\gamma t}$$

where $\bar{X}_t = t^{-1}\sum_{s=1}^{t} X_{s+1/2}$ *is the ergodic average of the algorithm's sequence of leading states.*

This result shows that the EG and 1-EG algorithms share the same convergence rate guarantees, so we can safely drop one gradient calculation per iteration in the monotone case. The proof of the theorem is based on the following technical lemma which enables us to treat the different variants of the 1-EG method in a unified way.

Lemma 2. *Assume that* V *satisfies Assumption 3 (monotonicity). Suppose further that the sequence* $(X_t)_{t\in\mathbb{N}/2}$ *of points in* $\mathbb{R}^d$ *satisfies the following "quasi-descent" inequality with* $\mu_s, \lambda_s \geq 0$*:*

$$\|X_{s+1} - p\|^2 \leq \|X_s - p\|^2 - 2\lambda_s\langle V(X_{s+1/2}), X_{s+1/2} - p\rangle + \mu_s - \mu_{s+1} \tag{3}$$

[4] More precisely, Rakhlin and Sridharan [39, 40] use the term Optimistic Mirror Descent (OMD) in reference to the Mirror-Prox method of Nemirovski [33], itself a variant of (EG) with projections defined by means of a Bregman function; for a related treatment, see Nesterov [35] and Juditsky et al. [23].

[5] To see this, note that the difference between two consecutive intermediate steps $X_{t-1/2}$ and $X_{t+1/2}$ can be written as $X_{t+1/2} = \Pi_{\mathcal{X}}(X_{t-1/2} - (\gamma_{t-1} + \gamma_t)V_{t-1/2} + \gamma_{t-1}V_{t-3/2})$. Writing (OG) in the form presented above shows that (OG) can also be viewed as a single-call variant of the FBF method of Tseng [44].

for all $p \in \mathcal{X}_R$ and all $s \in \{1, \dots, t\}$. Then,

$$\mathrm{Err}_R\left(\frac{\sum_{s=1}^t \lambda_s X_{s+1/2}}{\sum_{s=1}^t \lambda_s}\right) \leq \frac{R^2 + \mu_1}{2\sum_{s=1}^t \lambda_s}.$$

Remark 1. For Examples 1 and 2 it is possible to state both Theorem 1 and Lemma 2 with more adapted measures. We refer the readers to the supplement for more details.

The use of Lemma 2 is tailored to time-averaged sequences like $\bar{X}_t$, and relies on establishing a suitable "quasi-descent inequality" of the form (3) for the iterates of 1-EG. Doing this requires in turn a careful comparison of successive iterates of the algorithm via the Lipschitz continuity assumption for V; we defer the precise treatment of this argument to the paper's supplement.

On the other hand, because the role of averaging is essential in this argument, the convergence of the algorithm's last iterate requires significantly different techniques. To the best of our knowledge, there are no comparable convergence rate guarantees for X_t under Assumptions 1–3; however, if Assumption 3 is strengthened to Assumption 3(s), the convergence of X_t to the (necessarily unique) solution of (VI) occurs at a geometric rate. For completeness, we state here a consolidated version of the geometric convergence results of Malitsky [26], Gidel et al. [19], and Mokhtari et al. [32].

Theorem 2. *Assume that V satisfies Assumptions 1, 2 and 3(s), and let $x^\star$ denote the (necessarily unique) solution of (VI). If a 1-EG algorithm is run with a sufficiently small step-size γ, the generated sequence X_t converges to $x^\star$ at a rate of $\|X_t - x^\star\| = \mathcal{O}(\exp(-\rho\, t))$ for some $\rho > 0$.*

4.2 Local convergence

We continue by presenting a local convergence result for deterministic, *non-monotone* problems. To state it, we will employ the following notion of regularity in lieu of Assumptions 1–3 and 3(s).

Definition 3. We say that $x^\star$ is a *regular solution* of (VI) if V is C^1-smooth in a neighborhood of $x^\star$ and the Jacobian $\mathrm{Jac}_V(x^\star)$ is positive-definite along rays emanating from $x^\star$, i.e.,

$$z^\top \mathrm{Jac}_V(x^\star) z \equiv \sum_{i,j=1}^d z_i \frac{\partial V_i}{\partial x_j}(x^\star) z_j > 0 \tag{4}$$

for all $z \in \mathbb{R}^d \setminus \{0\}$ that are tangent to $\mathcal{X}$ at $x^\star$.

This notion of regularity is an extension of similar conditions that have been employed in the local analysis of loss minimization and saddle-point problems. More precisely, if $V = \nabla f$ for some loss function f, this definition is equivalent to positive-definiteness of the Hessian along qualified constraints [5, Chap. 3.2]. As for saddle-point problems and smooth games, variants of this condition can be found in several different sources, see e.g., [17, 25, 30, 41, 42] and references therein.

Under this condition, we obtain the following local geometric convergence result for 1-EG methods.

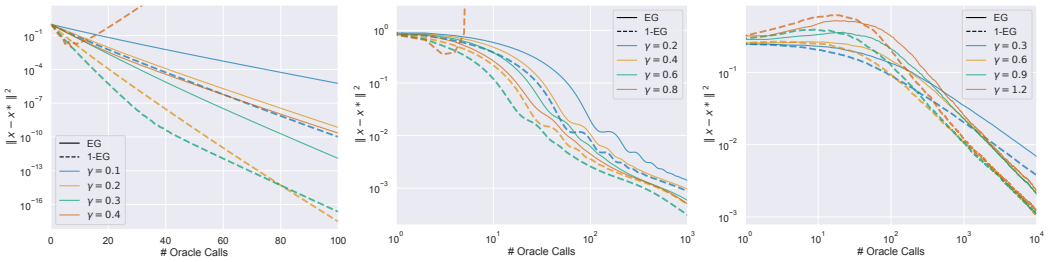
Theorem 4. *Let $x^\star$ be a regular solution of (VI). If a 1-EG method is run with perfect oracle feedback and is initialized sufficiently close to $x^\star$ with a sufficiently small constant step-size,we have $\|X_t - x^\star\| = \mathcal{O}(\exp(-\rho\, t))$ for some $\rho > 0$.*

The proof of this theorem relies on showing that (*i*) V essentially behaves like a smooth, strongly monotone operator close to $x^\star$; and (*ii*) if the method is initialized in a small enough neighborhood of $x^\star$, it will remain in said neighborhood for all t. As a result, Theorem 4 essentially follows by "localizing" Theorem 2 to this neighborhood.

As a preamble to our stochastic analysis in the next section, we should state here that, albeit straightforward, the proof strategy outlined above breaks down if we have access to V only via a *stochastic* oracle. In this case, a single "bad" realization of the feedback noise Z_t could drive the process away from the attraction region of any local solution of (VI). For this reason, the stochastic analysis requires significantly different tools and techniques and is considerably more intricate.

5 Stochastic analysis

We now present our analysis for stochastic variational inequalities with oracle feedback of the form (1). For concreteness, given that the PEG variant of the 1-EG method employs the most straightforward



(a) Strongly monotone [$\epsilon_1 = 1$, $\epsilon_2 = 0$], deterministic, last iterate **(b)** Monotone [$\epsilon_1 = 0$, $\epsilon_2 = 1$], deterministic, ergodic averaging **(c)** Non-monotone [$\epsilon_1 = 1$, $\epsilon_2 = -1$], iid $Z \sim \mathcal{N}(0, .01)$, last iterate ($b = 15$)

Figure 1: Illustration of the performance of EG and 1-EG in the (a priori non-monotone) saddle-point problem

$$\mathcal{L}(\theta, \phi) = 2\epsilon_1 \theta^\top A_1 \theta + \epsilon_2 \big(\theta^\top A_2 \theta\big)^2 - 2\epsilon_1 \phi^\top B_1 \phi - \epsilon_2 \big(\phi^\top B_2 \phi\big)^2 + 4\theta^\top C \phi$$

on the full unconstrained space $\mathcal{X} = \mathbb{R}^d = \mathbb{R}^{d_1 \times d_2}$ with $d_1 = d_2 = 1000$ and $A_1, B_1, A_2, B_2 \succ 0$. We choose three situations representative of the settings considered in the paper: (*a*) linear convergence of the last iterate of the deterministic methods in strongly monotone problems; (*b*) the $\mathcal{O}(1/t)$ convergence of the ergodic average in monotone, deterministic problems; and (*c*) the $\mathcal{O}(1/t)$ local convergence rate of the method's last iterate in stochastic, *non-monotone* problems. For (*a*) and (*b*), the origin is the unique solution of (VI), and for (*c*) it is a regular solution thereof. We observe that 1-EG consistently outperforms EG in terms of oracle calls for a fixed step-size, and the observed rates are consistent with the rates reported in Table 1.

proxy mechanism, we will focus on this variant throughout; for the other variants, the proofs and corresponding explicit expressions follow from the same rationale (as in the case of Theorem 1).

5.1 Global convergence

As we mentioned in the introduction, under Assumptions 1–3, Cui and Shanbhag [14] and Gidel et al. [19] showed that 1-EG methods attain a $\mathcal{O}(1/\sqrt{t})$ ergodic convergence rate. By strengthening Assumption 3 to Assumption 3(s), we show that this result can be augmented in two synergistic ways: under Assumptions 1, 2 and 3(s), both the last iterate and the ergodic average of 1-EG achieve a $\mathcal{O}(1/t)$ convergence rate.

Theorem 5. *Suppose that V satisfies Assumptions 1, 2 and 3(s), and assume that (PEG) is run with stochastic oracle feedback of the form (1) and a step-size of the form $\gamma_t = \gamma/(t+b)$ for some $\gamma > 1/\alpha$ and $b \geq 4\beta\gamma$. Then, the generated sequence of the algorithm's base states satisfies*

$$\mathbb{E}[\|X_t - x^\star\|^2] \leq \frac{6\gamma^2\sigma^2}{\alpha\gamma - 1}\frac{1}{t} + o\left(\frac{1}{t}\right), \tag{5}$$

while its ergodic average $\bar{X}_t = t^{-1}\sum_{s=1}^t X_s$ enjoys the bound

$$\mathbb{E}[\|\bar{X}_t - x^\star\|^2] \leq \frac{6\gamma^2\sigma^2}{\alpha\gamma - 1}\frac{\log t}{t} + o\left(\frac{\log t}{t}\right). \tag{6}$$

Regarding our proof strategy for the last iterate of the process, we can no longer rely either on a contraction argument or the averaging mechanism that yields the $\mathcal{O}(1/\sqrt{t})$ ergodic convergence rate. Instead, we show in the appendix that X_t is (stochastically) quasi-Fejér in the sense of [12, 13]; then, leveraging the method's specific step-size, we employ successive numerical sequence estimates to control the summability error and obtain the $\mathcal{O}(1/t)$ rate.

5.2 Local convergence

We proceed to examine the convergence of the method in the stochastic, *non-monotone* case. Our main result in this regard is the following.

Theorem 6. *Let $x^\star$ be a regular solution of (VI) and fix a tolerance level $\delta > 0$. Suppose further that (PEG) is run with stochastic oracle feedback of the form (1) and a variable step-size of the form $\gamma_t = \gamma/(t+b)$ for some $\gamma > 1/\alpha$ and large enough b. Then:*

(*a*) *There are neighborhoods U and U_1 of $x^\star$ in $\mathcal{X}$ such that, if $X_{1/2} \in U, X_1 \in U_1$, the event*

$$E_\infty = \{X_{t+1/2} \in U \text{ for all } t = 1, 2, \dots\}$$

occurs with probability at least $1 - \delta$.

(*b*) *Conditioning on the above, we have:*

$$\mathbb{E}[\|X_t - x^\star\|^2 \mid E_\infty] \leq \frac{4\gamma^2(M^2 + \sigma^2)}{(\alpha\gamma - 1)(1 - \delta)} \frac{1}{t} + o\left(\frac{1}{t}\right),$$

where $M = \sup_{x \in U} \|V(x)\| < \infty$ and $\alpha = \inf_{x \in U} \langle V(x), x - x^\star \rangle / \|x - x^\star\|^2 > 0$.

The finiteness of M and the positivity of α are both consequences of the regularity of $x^\star$ and their values only depend on the size of the neighborhood U. Taking a larger U would increase the algorithm's certified initialization basin but it would also negatively impact its convergence rate (since M would increase while α would decrease). Likewise, the neighborhood U_1 only depends on the size of U and, as we explain in the appendix, it suffices to take U_1 to be "one fourth" of U.

From the above, it becomes clear that the situation is significantly more involved than the corresponding deterministic analysis. This is also reflected in the proof of Theorem 6 which requires completely new techniques, well beyond the straightforward localization scheme underlying Theorem 4. More precisely, a key step in the proof (which we detail in the appendix) is to show that the iterates of the method remain close to $x^\star$ for all t with arbitrarily high probability. In turn, this requires showing that the probability of getting a string of "bad" noise realizations of arbitrary length is controllably small. Even then however, the global analysis *still* cannot be localized because conditioning changes the probability law under which the oracle noise is unbiased. Accounting for this conditional bias requires a surprisingly delicate probabilistic argument which we also detail in the supplement.

6 Concluding remarks

Our aim in this paper was to provide a synthetic view of single-call surrogates to the Extra-Gradient algorithm, and to establish optimal convergence rates in a range of different settings – deterministic, stochastic, and/or non-monotone. Several interesting avenues open up as a result, from extending the theory to more general Bregman proximal settings, to developing an adaptive version as in the recent work [2] for two-call methods. We defer these research directions to future work.

Acknowledgments

This work benefited from financial support by MIAI Grenoble Alpes (Multidisciplinary Institute in Artificial Intelligence). P. Mertikopoulos was partially supported by the French National Research Agency (ANR) grant ORACLESS (ANR–16–CE33–0004–01) and the EU COST Action CA16228 "European Network for Game Theory" (GAMENET).

References

[1] Adolphs, Leonard, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann. 2019. Local saddle point optimization: a curvature exploitation approach. *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

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

[3] Balduzzi, David, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, Thore Graepel. 2018. The mechanics of n-player differentiable games. *ICML '18: Proceedings of the 35th International Conference on Machine Learning*.

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

[5] Bertsekas, Dimitri P. 1997. Nonlinear programming. *Journal of the Operational Research Society* **48**(3) 334–334.

[6] Boţ, Radu Ioan, Panayotis Mertikopoulos, Mathias Staudigl, Phan Tu Vuong. 2019. Forward-backward-forward methods with variance reduction for stochastic variational inequalities. https://arxiv.org/abs/1902.03355.

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

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

[9] Chavdarova, Tatjana, Gauthier Gidel, François Fleuret, Simon Lacoste-Julien. 2019. Reducing noise in GAN training with variance reduced extragradient. https://arxiv.org/abs/1904.08598.

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

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

[12] Combettes, Patrick L. 2001. Quasi-Fejérian analysis of some optimization algorithms. Dan Butnariu, Yair Censor, Simeon Reich, eds., *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*. Elsevier, New York, NY, USA, 115–152.

[13] Combettes, Patrick L., Jean-Christophe Pesquet. 2015. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization* **25**(2) 1221–1248.

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

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

[16] Daskalakis, Constantinos, Ioannis Panageas. 2018. The limit points of (optimistic) gradient descent in min-max optimization. *NIPS'18: Proceedings of the 31st International Conference on Neural Information Processing Systems*.

[17] Facchinei, Francisco, Christian Kanzow. 2007. Generalized Nash equilibrium problems. *4OR* **5**(3) 173–210.

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

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

[20] Gidel, Gauthier, Reyhane Askari Hemmat, Mohammad Pezehski, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, Ioannis Mitliagkas. 2019. Negative momentum for improved game dynamics. *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

[21] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. 2014. Generative adversarial nets. *NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems*.

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

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

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

[25] Liang, Tengyuan, James Stokes. 2019. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

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

[27] Malitsky, Yura. 2018. Golden ratio algorithms for variational inequalities. https://arxiv.org/abs/1803.08832.

[28] Mazumdar, Eric V, Michael I Jordan, S Shankar Sastry. 2019. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. https://arxiv.org/abs/1901.00838.

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

[30] Mertikopoulos, Panayotis, Zhengyuan Zhou. 2019. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming* **173**(1-2) 465–507.

[31] Mokhtari, Aryan, Asuman Ozdaglar, Sarath Pattathil. 2019. 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.

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

[33] Nemirovski, Arkadi Semen. 2004. 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.

[34] Nesterov, Yurii. 2004. *Introductory Lectures on Convex Optimization: A Basic Course*. No. 87 in Applied Optimization, Kluwer Academic Publishers.

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

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

[37] Peng, Wei, Yu-Hong Dai, Hui Zhang, Lizhi Cheng. 2019. Training GANs with centripetal acceleration. https://arxiv.org/abs/1902.08949.

[38] Popov, Leonid Denisovich. 1980. 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.

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

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

[41] Ratliff, Lillian J, Samuel A Burden, S Shankar Sastry. 2013. Characterization and computation of local nash equilibria in continuous games. *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 917–924.

[42] Rosen, J. B. 1965. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica* **33**(3) 520–534.

[43] Tseng, Paul. 1995. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics* **60**(1-2) 237–252.

[44] Tseng, Paul. 2000. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization* **38**(2) 431–446.

[45] Yadav, Abhay, Sohil Shah, Zheng Xu, David Jacobs, Tom Goldstein. 2018. Stabilizing adversarial nets with prediction methods. *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*.

A Technical lemmas

Lemma A.1. *Let $x, y \in \mathbb{R}^d$ and $\mathcal{C} \subseteq \mathbb{R}^d$ be a closed convex set. We set $x^+ \coloneqq \Pi_{\mathcal{C}}(x - y)$. For all $p \in \mathcal{C}$, we have*

$$\|x^+ - p\|^2 \leq \|x - p\|^2 - 2\langle y, x^+ - p\rangle - \|x^+ - x\|^2.$$

Proof. Since $p \in \mathcal{C}$, we have the following property $\langle x^+ - (x - y), x^+ - p\rangle \leq 0$, leading to

$$\begin{aligned}
\|x^+ - p\|^2 &= \|x^+ - x + x - p\|^2 \\
&= \|x - p\|^2 + 2\langle x^+ - x, x - p\rangle + \|x^+ - x\|^2 \\
&= \|x - p\|^2 + 2\langle x^+ - x, x^+ - p\rangle - \|x^+ - x\|^2 \\
&\leq \|x - p\|^2 - 2\langle y, x^+ - p\rangle - \|x^+ - x\|^2.
\end{aligned}$$

□

Lemma A.2. *Let $x, y_1, y_2 \in \mathbb{R}^d$ and $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathbb{R}^d$ be two closed convex sets. We set $x_1^+ \coloneqq \Pi_{\mathcal{C}_1}(x - y_1)$ and $x_2^+ \coloneqq \Pi_{\mathcal{C}_2}(x - y_2)$.*

(a) *If $\mathcal{C}_2 = \mathbb{R}^d$, for all $p \in \mathbb{R}^d$, it holds*

$$\|x_2^+ - p\|^2 = \|x - p\|^2 - 2\langle y_2, x_1^+ - p\rangle + \|x_2^+ - x_1^+\|^2 - \|x_1^+ - x\|^2.$$

(b) *If $\mathcal{C}_2 \subseteq \mathcal{C}_1$, for all $p \in \mathcal{C}_2$, it holds*

$$\begin{aligned}
\|x_2^+ - p\|^2 &\leq \|x - p\|^2 - 2\langle y_2, x_1^+ - p\rangle + 2\langle y_2 - y_1, x_1^+ - x_2^+\rangle \\
&\quad - \|x_2^+ - x_1^+\|^2 - \|x_1^+ - x\|^2 \\
&\leq \|x - p\|^2 - 2\langle y_2, x_1^+ - p\rangle + \|y_2 - y_1\|^2 - \|x_1^+ - x\|^2. \qquad \text{(A.1)}
\end{aligned}$$

Proof. (a) We develop

$$\begin{aligned}
\|x_2^+ - p\|^2 &= \|x_2^+ - x_1^+ + x_1^+ - x + x - p\|^2 \\
&= \|x_2^+ - x_1^+\|^2 + \|x_1^+ - x\|^2 + \|x - p\|^2 \\
&\quad + 2\langle x_2^+ - x_1^+, x_1^+ - p\rangle + 2\langle x^+ - x, x - p\rangle \\
&= \|x_2^+ - x_1^+\|^2 - \|x_1^+ - x\|^2 + \|x - p\|^2 \\
&\quad + 2\langle x_2^+ - x_1^+, x_1^+ - p\rangle + 2\langle x_1^+ - x, x_1^+ - p\rangle \\
&= \|x - p\|^2 - 2\langle y_2, x_1^+ - p\rangle + \|x_2^+ - x_1^+\|^2 - \|x_1^+ - x\|^2,
\end{aligned}$$

where in the last line we use $x_2^+ - x = -y_2$ since $\mathcal{C}_2 = \mathbb{R}^d$.

(b) With $x_2^+ \in \mathcal{C}_2 \subseteq \mathcal{C}_1$, we can apply Lemma A.1 to $(x, y, x^+, p, \mathcal{C}) \leftarrow (x, y_2, x_2^+, p, \mathcal{C}_2)$ and $(x, y, x^+, p, \mathcal{C}) \leftarrow (x, y_1, x_1^+, x_2^+, \mathcal{C}_1)$, which yields

$$\|x_2^+ - p\|^2 \leq \|x - p\|^2 - 2\langle y_2, x_2^+ - p\rangle - \|x_2^+ - x\|^2, \qquad \text{(A.2)}$$

$$\|x_1^+ - x_2^+\|^2 \leq \|x - x_2^+\|^2 - 2\langle y_1, x_1^+ - x_2^+\rangle - \|x_1^+ - x\|^2. \qquad \text{(A.3)}$$

By summing (A.2) and (A.3), we readily get the first inequality of (A.1). We conclude with help of Young's inequality $2\langle y_2 - y_1, x_1^+ - x_2^+\rangle \leq \|y_2 - y_1\|^2 + \|x_1^+ - x_2^+\|^2$. □

Lemma A.3 (Chung [11, Lemma 1])**.** *Let $(a_t)_{t\in\mathbb{N}}$ be a sequence of real numbers and $b, t_0 \in \mathbb{N}$ such that for all $t \geq t_0$,*

$$a_{t+1} \leq \left(1 - \frac{q}{t+b}\right) a_t + \frac{q'}{(t+b)^2}, \qquad \text{(A.4)}$$

where $q > 1$ and $q' > 0$. Then,

$$a_t \leq \frac{q'}{q-1}\frac{1}{t} + o\left(\frac{1}{t}\right). \qquad \text{(A.5)}$$

Proof. For the sake of completeness, we provide a basic proof for the above lemma (which is a direct corollary of Chung [11, Lemma 1]). Let $q > 1$ and $k \in \mathbb{N}$, we have

$$\frac{1}{k+1} - \left(1 - \frac{q}{k}\right)\frac{1}{k} = \frac{q}{k^2} - \left(\frac{1}{k} - \frac{1}{k+1}\right) = \frac{q-1}{k^2} + \frac{1}{k^2(k+1)}.$$

This shows that for any $q' > 0$

$$\frac{q'}{q-1}\left(\frac{1}{k+1} - \left(1 - \frac{q}{k}\right)\frac{1}{k}\right) = \frac{q'}{k^2} + \frac{q'}{k^2(k+1)(q-1)} \geq \frac{q'}{k^2}. \tag{A.6}$$

By substituting $k \leftarrow t + b$, (A.4) combined with (A.6) yields

$$a_{t+1} - \frac{q'}{q-1}\frac{1}{t+b+1} \leq \left(1 - \frac{q}{t+b}\right)\left(a_t - \frac{q'}{q-1}\frac{1}{t+b}\right). \tag{A.7}$$

Let us define $a'_t \coloneqq a_t - q'/((q-1)(t+b))$. (A.7) becomes

$$a'_{t+1} \leq \left(1 - \frac{q}{t+b}\right) a'_t. \tag{A.8}$$

This inequality holds for all $t \geq t_0$. Then, either:
- a'_t becomes non-positive for some $t > t_1 = \max(t_0, \lfloor q \rfloor - b)$, and (A.8) implies that this is also the case for all subsequent t, which leads to

$$a_t \leq \frac{q'}{q-1}\frac{1}{t+b}.$$

- or a'_t is positive for all $t > t_1$ and we get

$$0 < a'_t \leq a'_{t_1} \prod_{s=t_1}^{t-1}\left(1 - \frac{q}{s+b}\right) = \mathcal{O}\left(\frac{1}{t^q}\right) = o\left(\frac{1}{t}\right).$$

In both cases, (A.5) is verified. □

Lemma A.4. *Let $x^\star$ be a regular solution of (VI). Then, there exists constants $r, \alpha, \beta > 0$ such that V is β-Lipschitz continuous on $\mathcal{K} \coloneqq \mathbb{B}_r(x^\star)$ and $\langle V(x), x - x^\star\rangle \geq \alpha\|x - x^\star\|^2$ for all $x \in U \coloneqq \mathcal{X} \cap \mathcal{K}$.*

Proof. The Lipschitz continuity is straightforward: a C^1-smooth operator is necessarily locally Lipschitz and thus Lipshitz on every compact. The proof consists in establishing the existence of α. To this end, we consider the following function:

$$\begin{array}{rccl} \phi\colon & \mathbb{R}^{d\times d} & \longrightarrow & \mathbb{R} \\ & G & \longmapsto & \min_{z \in \mathrm{TC}_{\mathcal{X}}(x^\star), \|z\|=1} z^\top G z \end{array}$$

where $\mathrm{TC}_{\mathcal{X}}(x^\star)$ denotes the tangent cone to $\mathcal{X}$ at $x^\star$. The function ϕ is concave as it is defined as a pointwise minimum over a set of linear functions. This in turn implies the continuity ϕ because every concave function is continous on the interior of its effective domain. The solution $x^\star$ being regular, we have $\phi(\mathrm{Jac}_V(x^\star)) > 0$. Combined with the continuity of Jac_V in a neighborhood of $x^\star$, we deduce the existence of $r, \alpha > 0$ such that $\phi(\mathrm{Jac}_V(x)) \geq \alpha$ for all $x \in \mathcal{K} = \mathbb{B}_r(x^\star)$. Now let $x \in U = \mathcal{X} \cap \mathcal{K}$. It holds:

$$V(x) - V(x^\star) = \left(\int_0^1 \mathrm{Jac}_V(x^\star + \lambda(x - x^\star))\, d\lambda\right)(x - x^\star).$$

Consequently, writing $z = x - x^\star \in \mathrm{TC}_{\mathcal{X}}(x^\star)$, $x'_\lambda = x^\star + \lambda(x - x^\star) \in \mathcal{K}$, we have

$$\begin{aligned} \langle V(x) - V(x^\star), x - x^\star\rangle &= z^\top \left(\int_0^1 \mathrm{Jac}_V(x'_\lambda)\, d\lambda\right) z \\ &\geq \left(\int \phi(\mathrm{Jac}_V(x'_\lambda))\, d\lambda\right)\|z\|^2 \geq \alpha\|z\|^2 = \alpha\|x - x^\star\|^2. \end{aligned}$$

Finally, since $x^\star$ is a solution of (VI), we have $\langle V(x^\star), x - x^\star\rangle \geq 0$ and

$$\langle V(x), x - x^\star\rangle \geq \langle V(x) - V(x^\star), x - x^\star\rangle \geq \alpha\|x - x^\star\|^2.$$

This ends the proof. □

B Proofs for the deterministic setting

B.1 Proof of Lemma 2

In the definition of Err_R, instead of taking $\mathcal{X}_R = \mathcal{X} \cap \mathbb{B}_R(0)$ we consider $\mathcal{X}_R = \mathcal{X} \cap \mathbb{B}_R(X_1)$. Summing (3) over s and rearranging the term leads to

$$\sum_{s=1}^{t} 2\lambda_s \langle V(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - p\rangle \leq \|X_1 - p\|^2 - \|X_{t+1} - p\|^2 + \mu_1 - \mu_{t+1} \leq \|X_1 - p\|^2 + \mu_1.$$

For any $p \in \mathcal{X}_R$, we have $\|X_1 - p\|^2 \leq R^2$, and by monoticity of V,

$$\langle V(p), X_{s+\frac{1}{2}} - p\rangle \leq \langle V(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - p\rangle.$$

In other words, for all $p \in \mathcal{X}_R$,

$$2\sum_{s=1}^{t} \lambda_s \langle V(p), X_{s+\frac{1}{2}} - p\rangle \leq R^2 + \mu_1. \tag{B.1}$$

Dividing the two sides of (B.1) by $2\sum_{s=1}^{t} \lambda_s$ and maximizing over $p \in \mathcal{X}_R$ leads to the desired result.

B.2 Proof of Theorem 1

To facilitate analysis and presentation of our results, (PEG) and (OG) are initialized with random $X_{\frac{1}{2}}$ and X_1 in $\mathcal{X}$ while for (RG) we start with X_0 and $X_{\frac{1}{2}}$. We are constrained to have different initial states in (RG) due to its specific formulation.

The theorem is immediate from Lemma 2 if we know that (3) is verified by the generated iterates for some $(\lambda_t)_{t\in\mathbb{N}}, (\mu_t)_{t\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$. Below, we show it separately for PEG, OG and RG under Assumption 2 and with γ selected as per the theorem statement. Moreover, we have $(\lambda_t)_{t\in\mathbb{N}} \equiv \gamma$ and $\mu_1 \leq \|X_1 - X_{\frac{1}{2}}\|^2$ for all methods, hence the corresponding bound in our statement. The arguments used in the proof are inspired from [19, 26, 44] but we emphasize the relation between the analyses of these algorithms by putting forward the technical Lemma A.2.

Past Extra-Gradient (PEG). For $t \geq 1$, the second inequality of Lemma A.2 (b) applied to $(x, y_1, y_2, x_1^+, x_2^+, \mathcal{C}_1, \mathcal{C}_2) \leftarrow (X_t, \gamma V(X_{t-\frac{1}{2}}), \gamma V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}, X_{t+1}, \mathcal{X}, \mathcal{X})$ results in

$$\begin{aligned}
\|X_{t+1} - p\|^2 &\leq \|X_t - p\|^2 - 2\gamma\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle \\
&\quad + \gamma^2\|V(X_{t+\frac{1}{2}}) - V(X_{t-\frac{1}{2}})\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2 \\
&\leq \|X_t - p\|^2 - 2\gamma\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle \\
&\quad + \gamma^2\beta^2\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2
\end{aligned} \tag{B.2}$$

where we used the fact that V is β-Lipschitz continuous for the second inequality.

Now, let us use Young's inequality $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ to get

$$\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 \leq 2\|X_{t+\frac{1}{2}} - X_t\|^2 + 2\|X_t - X_{t-\frac{1}{2}}\|^2 \tag{B.3}$$

and the non-expansiveness of the projection to get for any $t \geq 2$,

$$\|X_t - X_{t-\frac{1}{2}}\|^2 \leq \|X_{t-1} - \gamma V(X_{t-\frac{1}{2}}) - X_{t-1} + \gamma V(X_{t-\frac{3}{2}})\|^2 \leq \gamma^2\beta^2\|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2. \tag{B.4}$$

Combining (B.3) and (B.4), we obtain

$$\begin{aligned}
\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 &\leq 2\|X_{t+\frac{1}{2}} - X_t\|^2 + 2\gamma^2\beta^2\|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2 \\
&\leq 2\|X_{t+\frac{1}{2}} - X_t\|^2 + \frac{1}{2}\|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2,
\end{aligned} \tag{B.5}$$

where we used the fact that $\gamma \leq 1/(2\beta)$ in the last inequality; and in order to display a telescopic term, we reformulate (B.5) as

$$\begin{aligned}\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 &= 2\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 \\ &\leq 4\|X_{t+\frac{1}{2}} - X_t\|^2 + \|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2. \end{aligned} \tag{B.6}$$

We now substitute (B.6) in (B.2) to get for all $t \geq 2$,

$$\begin{aligned}\|X_{t+1} - p\|^2 &\leq \|X_t - p\|^2 - 2\gamma\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle + (4\gamma^2\beta^2 - 1)\|X_{t+\frac{1}{2}} - X_t\|^2 \\ &\quad + \gamma^2\beta^2(\|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2) \\ &\leq \|X_t - p\|^2 - 2\gamma\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle \\ &\quad + \gamma^2\beta^2(\|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2),\end{aligned}$$

and thus (3) holds true for all $t \geq 2$ with $\lambda_t = \gamma$ and $\mu_t = \gamma^2\beta^2\|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2$.

Finally, for $t = 1$, we have

$$\begin{aligned}&\gamma^2\beta^2\|X_{\frac{3}{2}} - X_{\frac{1}{2}}\|^2 - \|X_{\frac{3}{2}} - X_1\|^2 \\ &\leq 4\gamma^2\beta^2\|X_{\frac{3}{2}} - X_1\|^2 + 4\gamma^2\beta^2\|X_1 - X_{\frac{1}{2}}\|^2 - \gamma^2\beta^2\|X_{\frac{3}{2}} - X_{\frac{1}{2}}\|^2 - \|X_{\frac{3}{2}} - X_1\|^2 \\ &\leq 4\gamma^2\beta^2\|X_1 - X_{\frac{1}{2}}\|^2 - \gamma^2\beta^2\|X_{\frac{3}{2}} - X_{\frac{1}{2}}\|^2,\end{aligned}$$

which, plugged into (B.2) gives

$$\begin{aligned}\|X_2 - p\|^2 &\leq \|X_1 - p\|^2 - 2\gamma\langle V(X_{\frac{3}{2}}), X_{\frac{3}{2}} - p\rangle + \gamma^2\beta^2\|X_{\frac{3}{2}} - X_{\frac{1}{2}}\|^2 - \|X_{\frac{3}{2}} - X_1\|^2 \\ &\leq \|X_1 - p\|^2 - 2\gamma\langle V(X_{\frac{3}{2}}), X_{\frac{3}{2}} - p\rangle + 4\gamma^2\beta^2\|X_1 - X_{\frac{1}{2}}\|^2 - \gamma^2\beta^2\|X_{\frac{3}{2}} - X_{\frac{1}{2}}\|^2 \end{aligned} \tag{B.7}$$

which also matches (3) for $t = 1$ with $\lambda_t = \gamma$, μ_2 as defined previously, and $\mu_1 = 4\gamma^2\beta^2\|X_1 - X_{\frac{1}{2}}\|^2 \leq \|X_1 - X_{\frac{1}{2}}\|^2$. Thus, Lemma 2 enables us to conclude the proof for Past Extra-Gradient (PEG).

Optimistic Gradient (OG). The update of OG with constant step-size γ can be written as

$$\begin{cases} X_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(X_t - \gamma V(X_{t-\frac{1}{2}})) \\ X_{t+1} = X_t - (X_t - X_{t+\frac{1}{2}} + \gamma V(X_{t+\frac{1}{2}}) - \gamma V(X_{t-\frac{1}{2}})) \end{cases}$$

In that form, we can use Lemma A.2 (a) with $(x, y_1, y_2, x_1^+, x_2^+, \mathcal{C}_1, \mathcal{C}_2) \leftarrow (X_t, \gamma V(X_{t-\frac{1}{2}}), X_t - X_{t+\frac{1}{2}} + \gamma V(X_{t+\frac{1}{2}}) - \gamma V(X_{t-\frac{1}{2}}), X_{t+\frac{1}{2}}, X_{t+1}, \mathcal{X}, \mathbb{R}^d)$ to get

$$\begin{aligned}\|X_{t+1} - p\|^2 &= \|X_t - p\|^2 + \|X_{t+1} - X_{t+\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2 \\ &\quad - 2\langle X_t - X_{t+\frac{1}{2}} + \gamma V(X_{t+\frac{1}{2}}) - \gamma V(X_{t-\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle. \end{aligned} \tag{B.8}$$

One the one hand, since $X_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(X_t - \gamma V(X_{t-\frac{1}{2}}))$ and $p \in \mathcal{X}$, we have

$$\langle X_{t+\frac{1}{2}} - (X_t - \gamma V(X_{t-\frac{1}{2}})), X_{t+\frac{1}{2}} - p\rangle \leq 0. \tag{B.9}$$

On the other other hand, by definition of X_{t+1} and the β-Lipschitz continuity of V,

$$\|X_{t+1} - X_{t+\frac{1}{2}}\|^2 = \gamma^2\|V(X_{t+\frac{1}{2}}) - V(X_{t-\frac{1}{2}})\|^2 \leq \gamma^2\beta^2\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2. \tag{B.10}$$

Then, applying the same arguments used to get (B.6), we can show that for all $t \geq 2$,

$$\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 \leq 4\|X_{t+\frac{1}{2}} - X_t\|^2 + \|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2. \tag{B.11}$$

Putting together (B.8), (B.9), (B.10), and (B.11), we obtain for $\gamma \leq 1/(2\beta)$ and for all $t \geq 2$,

$$\begin{aligned}&\|X_{t+1} - p\|^2 \\ &\leq \|X_t - p\|^2 - 2\gamma\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle + \gamma^2\beta^2\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2 \\ &\leq \|X_t - p\|^2 - 2\gamma\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle + \gamma^2\beta^2(\|X_{t-\frac{1}{2}} - X_{t-\frac{3}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2).\end{aligned}$$

Finally, since (B.7) is still true using the same argument as for PEG, (3) is satisfied by choosing the same $(\mu_t)_{t\in\mathbb{N}}$ and $(\lambda_t)_{t\in\mathbb{N}}$ as in the case of PEG; the same result thus holds for Optimistic Gradient (OG).

Reflected Gradient (RG). We recall the update rule of RG

$$\begin{cases} X_{t+\frac{1}{2}} = X_t - (X_{t-1} - X_t) \\ X_{t+1} = \Pi_{\mathcal{X}}(X_t - \gamma V(X_{t+\frac{1}{2}})). \end{cases}$$

As in the previous cases, we use Lemma A.2. Using the first inequality of Part (b) with $(x, y_1, y_2, x_1^+, x_2^+, \mathcal{C}_1, \mathcal{C}_2) \leftarrow (X_t, X_{t-1} - X_t, \gamma V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}}, X_{t+1}, \mathbb{R}^d, \mathcal{X})$, we get

$$\begin{aligned} \|X_{t+1} - p\|^2 \leq \|X_t - p\|^2 &+ 2\langle \gamma V(X_{t+\frac{1}{2}}) - (X_{t-1} - X_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle \\ &- 2\gamma \langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle - \|X_{t+1} - X_{t+\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2. \end{aligned} \tag{B.12}$$

As $X_t = \Pi_{\mathcal{X}}(X_{t-1} - \gamma V(X_{t-\frac{1}{2}}))$ and $X_{t-1}, X_{t+1} \in \mathcal{X}$, it follows

$$\langle X_t - (X_{t-1} - \gamma V(X_{t-\frac{1}{2}})), X_t - X_{t-1} \rangle \leq 0, \tag{B.13}$$

$$\langle X_t - (X_{t-1} - \gamma V(X_{t-\frac{1}{2}})), X_t - X_{t+1} \rangle \leq 0. \tag{B.14}$$

By summing (B.13) and (B.14) and rearranging the terms, we get

$$\langle X_t - X_{t-1}, X_{t+\frac{1}{2}} - X_{t+1} \rangle \leq -\langle \gamma V(X_{t-\frac{1}{2}}), X_{t+\frac{1}{2}} - X_{t+1} \rangle,$$

thus,

$$\begin{aligned} &2\langle \gamma V(X_{t+\frac{1}{2}}) - (X_{t-1} - X_t), X_{t+\frac{1}{2}} - X_{t+1} \rangle \\ &\quad \leq 2\langle \gamma V(X_{t+\frac{1}{2}}) - \gamma V(X_{t-\frac{1}{2}}), X_{t+\frac{1}{2}} - X_{t+1} \rangle \\ &\quad \leq 2\gamma\beta \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\| \|X_{t+\frac{1}{2}} - X_{t+1}\|. \end{aligned} \tag{B.15}$$

Combining (B.12) and (B.15), we get

$$\begin{aligned} \|X_{t+1} - p\|^2 \leq \|X_t - p\|^2 &+ 2\gamma\beta \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\| \|X_{t+\frac{1}{2}} - X_{t+1}\| \\ &- 2\gamma \langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle - \|X_{t+1} - X_{t+\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2. \end{aligned} \tag{B.16}$$

By using twice Young's inequality: i) $2\langle a, b \rangle \leq \varepsilon \|a\|^2 + (1/\varepsilon)\|b\|^2$ with $\varepsilon = 1/\sqrt{2}$; then ii) $\|a + b\|^2 \leq (1 + \varepsilon')\|a\|^2 + (1 + 1/\varepsilon')\|b\|^2$ with $\varepsilon' = 1 + \sqrt{2}$, we have

$$\begin{aligned} &2\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\| \|X_{t+\frac{1}{2}} - X_{t+1}\| \\ &\quad \leq \frac{1}{\sqrt{2}} \|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 + \sqrt{2} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 \\ &\quad \leq (1 + \sqrt{2}) \|X_{t+\frac{1}{2}} - X_t\|^2 + \|X_t - X_{t-\frac{1}{2}}\|^2 + \sqrt{2} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2. \end{aligned} \tag{B.17}$$

Substituting (B.17) into (B.16) yields

$$\begin{aligned} \|X_{t+1} - p\|^2 &\leq \|X_t - p\|^2 - 2\gamma \langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &\quad + ((1 + \sqrt{2})\gamma\beta - 1) \|X_{t+\frac{1}{2}} - X_t\|^2 \\ &\quad + \gamma\beta \|X_t - X_{t-\frac{1}{2}}\|^2 - (1 - \sqrt{2}\gamma\beta) \|X_{t+1} - X_{t+\frac{1}{2}}\|^2 \\ &\leq \|X_t - p\|^2 - 2\gamma \langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p \rangle \\ &\quad + \gamma\beta \|X_t - X_{t-\frac{1}{2}}\|^2 - \gamma\beta \|X_{t+1} - X_{t+\frac{1}{2}}\|^2, \end{aligned}$$

where in the last line, we used twice that $\gamma \leq 1/((1 + \sqrt{2})\beta)$. Once again, (3) is verified with the choice $\forall t \in \mathbb{N}, \mu_t = \gamma\beta \|X_t - X_{t-\frac{1}{2}}\|^2, \lambda_t = \gamma$ and the result thus holds for Reflected Gradient (RG). We also notice that $\mu_1 \leq \|X_1 - X_{\frac{1}{2}}\|^2$ since $\gamma\beta < 1$.

B.3 Lemma 2 with other suboptimality measures

Here we discuss how the statement of Lemma 2, and consequently also that of Theorem 1, can be adjusted to consider more adapted convergence measures in the cases of loss minimization and min-max optimization. The notations are those of Examples 1 and 2, and we write $\bar{x} = (\sum_{s=1}^t \lambda_s)^{-1} \sum_{s=1}^t \lambda_s X_{s+\frac{1}{2}}$.

Loss minimization. $V = \nabla f$ is monotone implies the convexity of f, so

$$\langle V(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - p\rangle = \langle \nabla f(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - p\rangle \geq f(X_{s+\frac{1}{2}}) - f(p).$$

With Jensen's inequality we get,

$$\left(\sum_{s=1}^{t}\lambda_s\right)^{-1}\sum_{s=1}^{t}\lambda_s\langle V(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - p\rangle \geq \left(\sum_{s=1}^{t}\lambda_s\right)^{-1}\sum_{s=1}^{t}\lambda_s f(X_{s+\frac{1}{2}}) - f(p) \geq f(\bar{x}) - f(p)$$

This is true for any $p \in \mathcal{X}$, and especially for $p \in \mathcal{X}^\star$. Let $R = \operatorname{dist}(x_1, \mathcal{X}^\star)$. By invoking (**??**), we conclude

$$f(\bar{x}) - \min f \leq \frac{R^2 + \mu_1}{2\sum_{s=1}^{t}\lambda_s}.$$

Min-max optimization. $V = (\nabla_\theta \mathcal{L}, -\nabla_\phi \mathcal{L})$ being monotone is equivalent to $\mathcal{L}$ being convex-concave. In such saddle-point problems, the quality of a candidate solution $\hat{x} = (\hat{\theta}, \hat{\phi})$ is often assessed via the *Nikaido–Isoda* function [**?**], defined here as

$$\mathrm{NI}(\hat{x}) = \sup_{\phi\in\Phi}\mathcal{L}(\hat{\theta}, \phi) - \inf_{\theta\in\Theta}\mathcal{L}(\theta, \hat{\phi}) \tag{NI}$$

provided of course that the right-hand side is well-posed. Its restricted variant NI_R can also be defined by analogy with the definition of Err_R.

Let us denote $X_{s+\frac{1}{2}} = (\theta_{s+\frac{1}{2}}, \phi_{s+\frac{1}{2}})$ and $p = (\theta, \phi)$. By convex-concavity of $\mathcal{L}$, it holds

$$\begin{aligned}\langle V(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - p\rangle &= \langle \nabla_\theta \mathcal{L}(\theta_{s+\frac{1}{2}}, \phi_{s+\frac{1}{2}}), \theta_{s+\frac{1}{2}} - \theta\rangle - \langle \nabla_\phi \mathcal{L}(\theta_{s+\frac{1}{2}}, \phi_{s+\frac{1}{2}}), \phi_{s+\frac{1}{2}} - \phi\rangle \\ &\geq \mathcal{L}(\theta_{s+\frac{1}{2}}, \phi_{s+\frac{1}{2}}) - \mathcal{L}(\theta, \phi_{s+\frac{1}{2}}) + \mathcal{L}(\theta_{s+\frac{1}{2}}, \phi) - \mathcal{L}(\theta_{s+\frac{1}{2}}, \phi_{s+\frac{1}{2}}) \\ &= \mathcal{L}(\theta_{s+\frac{1}{2}}, \phi) - \mathcal{L}(\theta, \phi_{s+\frac{1}{2}}).\end{aligned}$$

We can again apply Jensen's inequality to show that

$$\left(\sum_{s=1}^{t}\lambda_s\right)^{-1}\sum_{s=1}^{t}\lambda_s\langle V(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - p\rangle \geq \mathcal{L}(\bar{\theta}, \phi) - \mathcal{L}(\theta, \bar{\phi}),$$

where we write $\bar{x} = (\bar{\theta}, \bar{\phi})$. By (**??**) and definition of the Nikaido–Isoda function, maximizing over $(\theta, \phi) \in \mathcal{X} \cap \mathbb{B}_R(X_1)$ gives

$$\mathrm{NI}_R(\bar{x}) \leq \frac{R^2 + \mu_1}{2\sum_{s=1}^{t}\lambda_s}.$$

B.4 Proof of Theorem 4

Here we provide a quick proof of Theorem 4. We do not try to optimize the constants and better results could be derived by examining each algorithm carefully. Note that since RG can evaluate V at infeasible points, we need to strengthen condition (4) in Definition 3 to consider all z in the *tangent span* of $\mathcal{X}$, i.e., the subspace of $\mathbb{R}^d$ spanned by all possible displacement vectors of the form $z = x' - x$, $x, x' \in \mathcal{X}$.

In order to show a local geometric convergence rate we only need to show that by choosing sufficiently small constant step-size and initializing at points sufficiently close to $x^\star$, we ensure $X_t \in \mathcal{K}$ for all $t \in \mathbb{N}/2$ (where $\mathcal{K}$ is defined in Lemma A.4 and this is in view of Theorem 2). In fact, although Theorem 2 is stated for strongly monotone operators, by carefully examining its proof, it turns out that we only need $\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle \geq \alpha\|X_{t+\frac{1}{2}} - x^\star\|^2$ for some constant $\alpha > 0$ and all $t \in \mathbb{N}$. We thus proceed to show that $\forall t \in \mathbb{N}/2, X_t \in \mathcal{K}$. To do so, let us show that one can choose the initial points and γ so that $\forall t \in \mathbb{N}$, (*i*) $\|X_t - x^\star\|^2 \leq \frac{r^2}{4}$; (*ii*) $X_{t+\frac{1}{2}} \in \mathcal{K}$.

Part (*i*). It is proved in Appendix B.2 that the iterates of the 1-EG methods verify (3) under Assumption 2 (Lipschitz continuity) if γ is smaller than some constant. By Lemma A.4 we know that V is indeed Lipschitz continuous on the compact $\mathcal{K}$. Suppose that for all $s \in \mathbb{N}/2$, $s \leq t$, we have $X_s \in \mathcal{K}$, then it holds $\langle V(X_{s+\frac{1}{2}}), X_{s+\frac{1}{2}} - x^\star\rangle \geq 0$ for all $s \in \{1, ..., t-1\}$. This is true for PEG

and OG because $X_{s+\frac{1}{2}} \in \mathcal{X}$ and subsequently $X_{s+\frac{1}{2}} \in U = \mathcal{X} \cup \mathcal{K}$. For RG we did mention above that we need to relax the definition of a regular solution to consider all the $z \in \mathbb{R}^d$ and the statement of Lemma A.4 can also be modified accordingly. Using (3), we obtain[6]

$$\|X_t - x^\star\|^2 + \mu_t \leq \|X_1 - x^\star\|^2 + \mu_1.$$

for the three algorithms with $\mu_t \geq 0$. By imposing $X_{\frac{1}{2}} = X_1$ in PEG and OG, we get $\mu_1 = 0$. Similarly, we may impose $X_0 = X_{\frac{1}{2}}$ in RG, leading to $\mu_1 \leq \|X_1 - X_0\|^2 \leq \gamma^2 \|V(X_0)\|^2$. It is thus possible to choose the adequate initial points and γ such that $\|X_1 - x^\star\|^2 + \mu_1 \leq \frac{r^2}{4}$, which in turn guarantees $\|X_t - x^\star\|^2 \leq \frac{r^2}{4}$.

Part (*ii*). We now proceed to prove that we may choose γ sufficiently small such that if $\|X_t - x^\star\|^2 \leq \frac{r^2}{4}$ and $X_{t-\frac{1}{2}} \in \mathcal{K}$ then $X_{t+\frac{1}{2}} \in \mathcal{K}$. We notice that for the three algorithms, we have

$$\|X_{t+\frac{1}{2}} - X_t\|^2 \leq \gamma^2 \|V(X_{t-\frac{1}{2}})\|^2$$

by the non-expansiveness of the projection.[7] We define $M \coloneqq \sup_{x \in \mathcal{K}} \|V(x)\| < \infty$ where the finiteness of M comes from the continuity of V and the boundedness of $\mathcal{K}$. We choose $\gamma \leq r/(2M)$ so that $\gamma^2 \|V(X_{t-\frac{1}{2}})\|^2 \leq \frac{r^2}{4}$ since $X_{t-\frac{1}{2}} \in \mathcal{K}$. Then, by Young's inequality, we get

$$\|X_{t+\frac{1}{2}} - x^\star\|^2 \leq 2\|X_{t+\frac{1}{2}} - X_t\|^2 + 2\|X_t - x^\star\|^2 \leq r^2.$$

In other words, $X_{t+\frac{1}{2}} \in \mathcal{K}$.

Conclusion. We first notice that the conditions on the initial points and the stepsize γ do not depend on the iteration. Thus, by simple induction we have that if we initialize the algorithm such that

$$\gamma \leq r/(2M) \quad \text{and} \quad \|X_1 - x^\star\|^2 + \mu_1 \leq \frac{r^2}{4},$$

then for all $t \in \mathbb{N}/2$, $X_t \in \mathcal{K}$, concluding the proof.

C Proofs for the stochastic setting

Let us focus in this section on the (PEG) algorithm:

$$\begin{aligned} X_{t+\frac{1}{2}} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_{t-\frac{1}{2}}) \\ X_{t+1} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V_{t+\frac{1}{2}}) \end{aligned} \tag{PEG}$$

Following Appendix B.2, we initialize the algorithm with random $X_{\frac{1}{2}}$ and X_1 in $\mathcal{X}$. Recall that $(\mathcal{F}_t)_{t \in \frac{\mathbb{N}}{2}}$ denotes the natural filtration associated with the sequence $(X_t)_{t \in \frac{\mathbb{N}}{2}}$. In the PEG algorithm, we have $\mathcal{F}_t = \mathcal{F}_{t+\frac{1}{2}}$ for all $t \in \mathbb{N}$ (thus $X_{t+\frac{1}{2}}$ is $\mathcal{F}_t$-measurable) so the zero-mean hypothesis (2a) can be written as $\mathbb{E}[Z_{t+\frac{1}{2}} \mid \mathcal{F}_t] = 0$.

C.1 Proof of Theorem 5

Last iterate convergence. As in the proof of Theorem 1, we first apply Lemma A.2 (b) with $(x, y_1, y_2, x_1^+, x_2^+, \mathcal{C}_1, \mathcal{C}_2) \leftarrow (X_t, \gamma_t V_{t-\frac{1}{2}}, \gamma_t V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}}, X_{t+1}, \mathcal{X}, \mathcal{X})$ and the solution $x^\star \in \mathcal{X}$ as a trial point to obtain

$$\begin{aligned} \|X_{t+1} - x^\star\|^2 \leq \|X_t - x^\star\|^2 &- 2\gamma_t \langle V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle \\ &+ \gamma_t^2 \|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2. \end{aligned} \tag{C.1}$$

The following holds true thanks to the law of total expectation,

$$\mathbb{E}[\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2]$$

[6] Please refer to the proof of Theorem 1 for the exact value of μ_t.

[7] In particular this also holds for RG since then $X_{t+\frac{1}{2}} - X_t = X_t - X_{t-1} = \Pi_{\mathcal{X}}(X_{t-1} - \gamma V(X_{t-\frac{1}{2}})) - \Pi_{\mathcal{X}}(X_{t-1})$.

$$
\begin{aligned}
&= \mathbb{E}[\|V(X_{t+\frac{1}{2}}) - V_{t-\frac{1}{2}}\|^2 + 2\langle Z_{t+\frac{1}{2}}, V(X_{t+\frac{1}{2}}) - V_{t-\frac{1}{2}}\rangle + \|Z_{t+\frac{1}{2}}\|^2] \\
&= \mathbb{E}[\|V(X_{t+\frac{1}{2}}) - V_{t-\frac{1}{2}}\|^2] + 2\,\mathbb{E}[\mathbb{E}[\langle Z_{t+\frac{1}{2}}, V(X_{t+\frac{1}{2}}) - V_{t-\frac{1}{2}}\rangle \mid \mathcal{F}_t]] + \mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2] \\
&= \mathbb{E}[\|V(X_{t+\frac{1}{2}}) - V_{t-\frac{1}{2}}\|^2] + \mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2]. \qquad \text{(C.2)}
\end{aligned}
$$

By Young's inequality, β-Lipschitz continuity of V, and non-expansiveness of the projection, we have

$$
\begin{aligned}
\|V(X_{t+\frac{1}{2}}) - V_{t-\frac{1}{2}}\|^2 &\le 2\|V(X_{t+\frac{1}{2}}) - V(X_{t-\frac{1}{2}})\|^2 + 2\|Z_{t-\frac{1}{2}}\|^2 \\
&\le 2\beta^2\|X_{t+\frac{1}{2}} - X_{t-\frac{1}{2}}\|^2 + 2\|Z_{t-\frac{1}{2}}\|^2 \\
&\le 4\beta^2\|X_{t+\frac{1}{2}} - X_t\|^2 + 4\beta^2\|X_t - X_{t-\frac{1}{2}}\|^2 + 2\|Z_{t-\frac{1}{2}}\|^2 \\
&\le 4\beta^2\|X_{t+\frac{1}{2}} - X_t\|^2 + 4\gamma_{t-1}^2\beta^2\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2 + 2\|Z_{t-\frac{1}{2}}\|^2. \qquad \text{(C.3)}
\end{aligned}
$$

Notice that the choice $b \ge 4\beta\gamma$ implies $8\gamma_t^2\beta^2 + 2\gamma_t\beta \le 1$, which in turn yields $8\gamma_t^2\beta^2 \le 1 - \alpha\gamma_t$. Combining (C.2) and (C.3), similarly to (B.6), we can thus show that

$$
\begin{aligned}
\mathbb{E}[\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2] &\le 8\beta^2\,\mathbb{E}[\|X_{t+\frac{1}{2}} - X_t\|^2] + 8\gamma_{t-1}^2\beta^2\,\mathbb{E}[\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2] \\
&\quad + 4\,\mathbb{E}[\|Z_{t-\frac{1}{2}}\|^2] + 2\,\mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2] - \mathbb{E}[\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2] \\
&\le 8\beta^2\,\mathbb{E}[\|X_{t+\frac{1}{2}} - X_t\|^2] + 6\sigma^2 \\
&\quad + \frac{\gamma_{t-1}^2}{\gamma_t^2}(1 - \alpha\gamma_t)\,\mathbb{E}[\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2] - \mathbb{E}[\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2], \qquad \text{(C.4)}
\end{aligned}
$$

where in the last line we also use $\mathbb{E}[\|Z_{t-\frac{1}{2}}\|^2] \le \sigma^2$, $\mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2] \le \sigma^2$.

We also have

$$
\mathbb{E}[\langle V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star\rangle] = \mathbb{E}[\mathbb{E}[\langle V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star\rangle \mid \mathcal{F}_t]] = \mathbb{E}[\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle]. \qquad \text{(C.5)}
$$

Since $x^\star$ is the unique solution of (VI), it follows $\langle V(x^\star), X_{t+\frac{1}{2}} - x^\star\rangle \ge 0$. Consequently, with strong monotonicity of V, we get

$$
\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle \ge \langle V(X_{t+\frac{1}{2}}) - V(x^\star), X_{t+\frac{1}{2}} - x^\star\rangle \ge \alpha\|X_{t+\frac{1}{2}} - x^\star\|^2.
$$

By Young's inequality

$$
\|X_t - x^\star\|^2 \le 2\|X_t - X_{t+\frac{1}{2}}\|^2 + 2\|X_{t+\frac{1}{2}} - x^\star\|^2,
$$

we can further write

$$
\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle \ge \frac{\alpha}{2}\|X_t - x^\star\|^2 - \alpha\|X_t - X_{t+\frac{1}{2}}\|^2. \qquad \text{(C.6)}
$$

Taking expectation over (C.1) and using (C.4), (C.5), (C.6) leads to

$$
\begin{aligned}
\mathbb{E}[\|X_{t+1} - x^\star\|^2] &\le \mathbb{E}[\|X_t - x^\star\|^2] - \alpha\gamma_t\,\mathbb{E}[\|X_t - x^\star\|^2] + 2\alpha\gamma_t\,\mathbb{E}[\|X_t - X_{t+\frac{1}{2}}\|^2] \\
&\quad + \gamma_{t-1}^2(1 - \alpha\gamma_t)\,\mathbb{E}[\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2] \\
&\quad + 8\gamma_t^2\beta^2\,\mathbb{E}[\|X_{t+\frac{1}{2}} - X_t\|^2] - \gamma_t^2\,\mathbb{E}[\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2] \\
&\quad + 6\gamma_t^2\sigma^2 - \mathbb{E}[\|X_{t+\frac{1}{2}} - X_t\|^2] \\
&= (1 - \alpha\gamma_t)(\mathbb{E}[\|X_t - x^\star\|^2] + \gamma_{t-1}^2\,\mathbb{E}[\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2]) \\
&\quad + 6\gamma_t^2\sigma^2 - \gamma_t^2\,\mathbb{E}[\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2] \\
&\quad + (8\gamma_t^2\beta^2 + 2\alpha\gamma_t - 1)\,\mathbb{E}[\|X_{t+\frac{1}{2}} - X_t\|^2]. \qquad \text{(C.7)}
\end{aligned}
$$

Using $8\gamma_t^2\beta^2 + 2\alpha\gamma_t - 1 \le 0$, (C.7) reduces to

$$
\begin{aligned}
&\mathbb{E}[\|X_{t+1} - x^\star\|^2] + \gamma_t^2\,\mathbb{E}[\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2] \\
&\qquad \le (1 - \alpha\gamma_t)(\mathbb{E}[\|X_t - x^\star\|^2] + \gamma_{t-1}^2\,\mathbb{E}[\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2]) + 6\gamma_t^2\sigma^2.
\end{aligned}
$$

We conclude by applying Lemma A.3 with $a_t \leftarrow \mathbb{E}[\|X_t - x^\star\|^2] + \gamma_{t-1}^2 \mathbb{E}[\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2]$, $q \leftarrow \alpha\gamma$, $q' \leftarrow 6\gamma^2\sigma^2$, and $t_0 \leftarrow 2$, which gives

$$\mathbb{E}[\|X_t - x^\star\|^2] + \gamma_{t-1}^2 \mathbb{E}[\|V_{t-\frac{1}{2}} - V_{t-\frac{3}{2}}\|^2] \leq \frac{6\gamma^2\sigma^2}{\alpha\gamma - 1}\frac{1}{t} + o\left(\frac{1}{t}\right).$$

The second term on the left-hand side (LHS) of the inequality is always positive, and (5) follows immediately.

Ergodic convergence. The convergence of $\bar{X}_t$ as shown in (6) can be deduce directly from above by using Jensen's inequality:

$$\mathbb{E}[\|\bar{X}_t - x^\star\|^2] \leq \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}[\|X_s - x^\star\|^2],$$

and then we bound the right-hand side (RHS) of the inequality by (5).

C.2 Proof of Theorem 6

We start by defining some important quantities that will be used in our proof. For any $T \geq 1$, we set

$$\begin{aligned} S_T &\coloneqq \sum_{t=1}^{T} 2\gamma_t \langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle, \\ R_T &\coloneqq \sum_{t=1}^{T} 2\gamma_t^2 (\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2), \\ Q_T &\coloneqq S_T^2 + R_T. \end{aligned}$$

Notice that S_T, R_T and Q_T are not $\mathcal{F}_T$-measurable but $\mathcal{F}_{T+1}$-measurable (due to the terms in $Z_{T+\frac{1}{2}}$ and $V_{T+\frac{1}{2}}$). For the sake of simplicity, we also write $\xi_{t+\frac{1}{2}} \coloneqq \langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle$ so that $S_T = \sum_{t=1}^{T} 2\gamma_t \xi_{t+\frac{1}{2}}$ and $\mathbb{E}[\xi_{t+\frac{1}{2}} \mid \mathcal{F}_t] = 0$.

Regarding the choice of U and U_1, we invoke Lemma A.4 to obtain the corresponding α, r and U. We then set $U_1 \coloneqq \mathcal{X} \cap \mathbb{B}_{r/4}(x^\star)$. Let us consider the following events for $T \geq 1$,

$$\begin{aligned} H_T &\coloneqq \left\{ \max_{1 \leq t \leq T} Q_t \leq \varepsilon \coloneqq \min\left(\frac{r^2}{8}, \frac{r^4}{16}\right) \right\}, \\ E_T &\coloneqq \left\{ \forall t \in \{1, ..., T\}, X_{t+\frac{1}{2}} \in U \right\}. \end{aligned}$$

We additionally define $Q_0 \coloneqq 2\gamma_1^2 \|V_{\frac{1}{2}}\|^2$, $H_0 \coloneqq \{Q_0 \leq \varepsilon\}$ and $H_{-1} \coloneqq E_0 \coloneqq \Omega$, where Ω denotes the whole sample space. It follows from the definitions that both $(H_T)_{T \geq -1}$ and $(E_T)_{T \geq 0}$ are decreasing sequences of events. Moreover, we have $H_T \in \mathcal{F}_{T+1}$ while $E_T \in \mathcal{F}_T$. Also notice that $E_\infty = \bigcap_{T \geq 0} E_T$.

In terms of notation, for an event $E \subseteq \Omega$, we denote by $\mathbb{1}_E$ its indicator function and E^c its complementary. For any pair of events $E, F \subseteq \Omega$, we denote by $E \setminus F$ the event "E and not F" i.e., $E \cap F^c$.

The proof of the theorem relies on the two following lemmas.

Lemma C.1. *For any $T \geq 0$, we have the inclusion $H_{T-1} \subseteq E_T$.*

Proof. We prove this result by induction.

Initialization: $H_{-1} \subseteq E_0$ is clear. To prove that we also have $H_0 \subseteq E_1$, we use Young's inequality to get

$$\|X_{\frac{3}{2}} - x^\star\|^2 \leq 2\|X_{\frac{3}{2}} - X_1\|^2 + 2\|X_1 - x^\star\|^2. \tag{C.8}$$

On the one hand, since $X_1 \in U_1$ by assumption, it holds $\|X_1 - x^\star\|^2 \leq \frac{r^2}{16}$. On the other hand,

$$2\|X_{\frac{3}{2}} - X_1\|^2 = 2\|\Pi_{\mathcal{X}}(X_1 - \gamma_1 V_{\frac{1}{2}}) - \Pi_{\mathcal{X}}(X_1)\|^2 \leq 2\gamma_1 \|V_{\frac{1}{2}}\|^2 = Q_0$$

For any realization in H_0, we have $2\gamma_1\|V_{\frac{1}{2}}\|^2 \leq \frac{r^2}{8}$; and so we can deduce from (C.8) that $\|X_{\frac{3}{2}} - x^\star\|^2 \leq \frac{r^2}{4} < r^2$. Since $X_{\frac{3}{2}} \in \mathcal{X}$, it follows that $X_{\frac{3}{2}} \in U$. This means that $H_0 \subseteq E_1$.

Inductive step: Suppose that $H_{T-1} \subseteq E_T$ holds for some $T \geq 1$. We would like to prove $H_T \subseteq E_{T+1}$. To do so, we show that $\|X_{T+1} - x^\star\|^2 \leq \frac{7}{16}r^2$ for any realization in H_T. Applying Lemma A.2 (b) as in (C.1) yields for all $t \in \{1, ..., T\}$,

$$\begin{aligned}
\|X_{t+1} - x^\star\|^2 &\leq \|X_t - x^\star\|^2 - 2\gamma_t\langle V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star\rangle \\
&\quad + \gamma_t^2\|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2 \\
&\leq \|X_t - x^\star\|^2 - 2\gamma_t\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle \\
&\quad - 2\gamma_t\langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star\rangle + 2\gamma_t^2(\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2) \\
&\leq \|X_t - x^\star\|^2 - 2\gamma_t\xi_{t+\frac{1}{2}} + 2\gamma_t^2(\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2), \qquad \text{(C.9)}
\end{aligned}$$

where in the last line we can use $\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle \geq 0$ since by induction hypothesis, $H_T \subseteq H_{T-1} \subseteq E_T$, which means for any realization in H_T, $X_{t+\frac{1}{2}} \in U$ for all $t \in \{1, ..., T\}$.

Summing (C.9) from $t = 1$ to T gives

$$\begin{aligned}
\|X_{T+1} - x^\star\|^2 &\leq \|X_1 - x^\star\|^2 - \sum_{t=1}^{T} 2\gamma_t\xi_{t+\frac{1}{2}} + \sum_{t=1}^{T} 2\gamma_t^2(\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2) \\
&= \|X_1 - x^\star\|^2 - S_T + R_T.
\end{aligned}$$

By definition of H_T, we have $S_T^2 \leq Q_T \leq \frac{r^4}{16}$ (so $|S_T| \leq \frac{r^2}{4}$) and $R_T \leq Q_T \leq \frac{r^2}{8}$. Using that $\|X_1 - x^\star\|^2 \leq \frac{r^2}{16}$ by assumption, it follows immediately that $\|X_{T+1} - x^\star\|^2 \leq \frac{7}{16}r^2$.

Finally, in order to bound $\|X_{T+\frac{3}{2}} - x^\star\|^2$, we again rely on Young's inequality:

$$\begin{aligned}
\|X_{T+\frac{3}{2}} - x^\star\|^2 &\leq 2\|X_{T+\frac{3}{2}} - X_{T+1}\|^2 + 2\|X_{T+1} - x^\star\|^2 \\
&\leq 2\gamma_{T+1}^2\|V_{T+\frac{1}{2}}\|^2 + 2\|X_{T+1} - x^\star\|^2. \qquad \text{(C.10)}
\end{aligned}$$

For any realization in H_T, we have that

$$\text{i)} \quad 2\gamma_{T+1}^2\|V_{T+\frac{1}{2}}\|^2 \leq 2\gamma_T^2\|V_{T+\frac{1}{2}}\|^2 \leq R_T \leq Q_T \leq \frac{r^2}{8};$$

$$\text{ii)} \quad 2\|X_{T+1} - x^\star\|^2 \leq \frac{7}{8}r^2.$$

Thus, (C.10) implies that $\|X_{T+\frac{3}{2}} - x^\star\|^2 \leq r^2$, and subsequently $X_{T+\frac{3}{2}} \in U$. As $H_T \subseteq E_T$ and $E_{T+1} = \{X_{T+\frac{3}{2}} \in U\} \cap E_T$, we have proven that $H_T \subseteq E_{T+1}$. □

Lemma C.2. *For $t \geq 1$, we have the following recurrence inequality*

$$\mathbb{E}[Q_t\,\mathbb{1}_{H_{t-1}}] \leq \mathbb{E}[Q_{t-1}\,\mathbb{1}_{H_{t-2}}] + \gamma_t^2\mathcal{M} - \varepsilon\,\mathbb{P}(H_{t-2} \setminus H_{t-1}), \qquad \text{(C.11)}$$

where $\mathcal{M} := 4M^2 + 4\sigma^2 + 4r^2\sigma^2$ and $\varepsilon := \min\left(\frac{r^2}{8}, \frac{r^4}{16}\right)$.

Moreover, if $t = 1$, the bound can be refined to

$$\mathbb{E}[Q_1\,\mathbb{1}_{H_0}] \leq \mathbb{E}[Q_0\,\mathbb{1}_{H_{-1}}] + \gamma_1^2(2M^2 + 2\sigma^2 + 4r^2\sigma^2) - \varepsilon\,\mathbb{P}(H_{-1} \setminus H_0). \qquad \text{(C.12)}$$

Proof. We decompose

$$\begin{aligned}
\mathbb{E}[Q_t\,\mathbb{1}_{H_{t-1}}] &= \mathbb{E}[(Q_t - Q_{t-1})\,\mathbb{1}_{H_{t-1}}] + \mathbb{E}[Q_{t-1}\,\mathbb{1}_{H_{t-1}}] \\
&= \mathbb{E}[(Q_t - Q_{t-1})\,\mathbb{1}_{H_{t-1}}] + \mathbb{E}[Q_{t-1}\,\mathbb{1}_{H_{t-2}}] - \mathbb{E}[Q_{t-1}\,\mathbb{1}_{H_{t-2}\setminus H_{t-1}}], \qquad \text{(C.13)}
\end{aligned}$$

where the second equality comes from the fact that as $H_{t-1} \subseteq H_{t-2}$, we have $H_{t-1} = H_{t-2} \setminus (H_{t-2} \setminus H_{t-1})$.

For $t \geq 2$, we write

$$\begin{aligned} Q_t &= S_t^2 + R_t \\ &= S_{t-1}^2 + 4\gamma_t \xi_{t+\frac{1}{2}} S_{t-1} + 4\gamma_t^2 \xi_{t+\frac{1}{2}}^2 + R_{t-1} + 2\gamma_t^2(\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2) \\ &= Q_{t-1} + 4\gamma_t \xi_{t+\frac{1}{2}} S_{t-1} + 4\gamma_t^2 \xi_{t+\frac{1}{2}}^2 + 2\gamma_t^2(\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2). \end{aligned} \tag{C.14}$$

Since S_{t-1} and H_{t-1} are $\mathcal{F}_t$-measurable, we get

$$\mathbb{E}[\xi_{t+\frac{1}{2}} S_{t-1} \mathbb{1}_{H_{t-1}}] = \mathbb{E}[\mathbb{E}[\xi_{t+\frac{1}{2}} \mid \mathcal{F}_t] S_{t-1} \mathbb{1}_{H_{t-1}}] = 0. \tag{C.15}$$

By Lemma C.1, $H_{t-1} \subseteq E_t$ which means that for any realization in H_{t-1}, we have $X_{t+\frac{1}{2}} \in U$. Therefore, $\|X_{t+\frac{1}{2}} - x^\star\|^2 \mathbb{1}_{H_{t-1}} \leq r^2 \mathbb{1}_{H_{t-1}}$ and consequently

$$\begin{aligned} \xi_{t+\frac{1}{2}}^2 \mathbb{1}_{H_{t-1}} &= \langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle^2 \mathbb{1}_{H_{t-1}} \\ &\leq \|Z_{t+\frac{1}{2}}\|^2 \|X_{t+\frac{1}{2}} - x^\star\|^2 \mathbb{1}_{H_{t-1}} \leq \|Z_{t+\frac{1}{2}}\|^2 r^2 \mathbb{1}_{H_{t-1}}. \end{aligned}$$

Using again that H_{t-1} is $\mathcal{F}_t$-measurable along with the boundedness of the variance of $Z_{t+\frac{1}{2}}$ (see Eq. (2b)), we get

$$\begin{aligned} \mathbb{E}[\xi_{t+\frac{1}{2}}^2 \mathbb{1}_{H_{t-1}}] &\leq r^2 \mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2 \mathbb{1}_{H_{t-1}}] = r^2 \mathbb{E}[\mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2 \mid \mathcal{F}_t] \mathbb{1}_{H_{t-1}}] \\ &\leq r^2 \mathbb{E}[\sigma^2 \mathbb{1}_{H_{t-1}}] = r^2 \sigma^2 \mathbb{P}[H_{t-1}] \leq r^2 \sigma^2. \end{aligned} \tag{C.16}$$

Applying once again the techniques above and relying on the boundedness of V (as for any realization in $H_{t-1} \subseteq E_t$ we have $X_{t+\frac{1}{2}} \in U$ and $M = \sup_{x \in U} V(x) < \infty$), we get

$$\begin{aligned} \mathbb{E}[\|V_{t+\frac{1}{2}}\|^2 \mathbb{1}_{H_{t-1}}] &= \mathbb{E}[(\|V(X_{t+\frac{1}{2}})\|^2 + 2\langle Z_{t+\frac{1}{2}}, V(X_{t+\frac{1}{2}})\rangle + \|Z_{t+\frac{1}{2}}\|^2) \mathbb{1}_{H_{t-1}}] \\ &= \mathbb{E}[\|V(X_{t+\frac{1}{2}})\|^2 \mathbb{1}_{H_{t-1}}] \\ &\quad + 2\,\mathbb{E}[\mathbb{E}[\langle Z_{t+\frac{1}{2}}, V(X_{t+\frac{1}{2}})\rangle \mid \mathcal{F}_t] \mathbb{1}_{H_{t-1}}] + \mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2 \mathbb{1}_{H_{t-1}}] \\ &= \mathbb{E}[\|V(X_{t+\frac{1}{2}})\|^2 \mathbb{1}_{H_{t-1}}] + 0 + \mathbb{E}[\mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2 \mid \mathcal{F}_t] \mathbb{1}_{H_{t-1}}] \\ &\leq M^2 + \sigma^2. \end{aligned} \tag{C.17}$$

Using that $H_{t-1} \subseteq H_{t-2}$ and repeating the arguments leading to (C.17), we have

$$\mathbb{E}[\|V_{t-\frac{1}{2}}\|^2 \mathbb{1}_{H_{t-1}}] \leq \mathbb{E}[\|V_{t-\frac{1}{2}}\|^2 \mathbb{1}_{H_{t-2}}] \leq M^2 + \sigma^2. \tag{C.18}$$

Combining (C.14), (C.15), (C.16), (C.17) and (C.18), we get

$$\mathbb{E}[(Q_t - Q_{t-1}) \mathbb{1}_{H_{t-1}}] \leq \gamma_t^2(4M^2 + 4\sigma^2 + 4r^2\sigma^2) = \gamma_t^2 \mathcal{M}. \tag{C.19}$$

For the last term on the RHS of (C.13), we get by definition that for any realization in $H_{t-2} \setminus H_{t-1}$, $Q_{t-1} > \varepsilon$ and thus

$$\mathbb{E}[Q_{t-1} \mathbb{1}_{H_{t-2} \setminus H_{t-1}}] \geq \varepsilon\, \mathbb{E}[\mathbb{1}_{H_{t-2} \setminus H_{t-1}}] = \varepsilon\, \mathbb{P}(H_{t-2} \setminus H_{t-1}). \tag{C.20}$$

Substituting (C.19) and (C.20) into (C.13) gives exactly (C.11).

The case $t = 1$ is proved similarly. In fact,

$$Q_1 - Q_0 = 4\gamma_1^2 \xi_{\frac{3}{2}}^2 + 2\gamma_1^2 \|V_{\frac{3}{2}}\|^2.$$

Consequently by using $H_0 \subseteq E_1$, we have

$$\mathbb{E}[(Q_1 - Q_0) \mathbb{1}_{H_0}] \leq \gamma_1^2(2M^2 + 2\sigma^2 + 4r^2\sigma^2).$$

By definition $H_{-1} \setminus H_0 = \{Q_0 > \varepsilon\}$, which shows (C.20) is equally true with $t = 1$. (C.12) can then be immediately deduced from (C.13). □

Proof of Theorem 6.

(a) We first show that by choosing b sufficiently large, we have $\mathbb{P}(H_T) \geq 1-\delta$ for all $T \geq -1$ (when $T=-1$, $H_{-1}=\Omega$). To do so, we will work on the complementary event ${H_T}^c = {H_{T-1}}^c \cup (H_{T-1} \setminus H_T)$ and prove that $\mathbb{P}({H_T}^c) \leq \delta$. We start by bounding $\mathbb{P}(H_{T-1} \setminus H_T)$,

$$\begin{aligned}
\varepsilon\, \mathbb{P}(H_{T-1} \setminus H_T) &= \varepsilon\, \mathbb{P}(\{Q_T > \varepsilon\} \cap H_{T-1}) \\
&= \mathbb{E}[\varepsilon\, \mathbb{1}_{\{Q_T > \varepsilon\} \cap H_{T-1}}] \\
&\leq \mathbb{E}[Q_T\, \mathbb{1}_{\{Q_T > \varepsilon\} \cap H_{T-1}}] \\
&\leq \mathbb{E}[Q_T\, \mathbb{1}_{H_{T-1}}]. \qquad \text{(C.21)}
\end{aligned}$$

The last line is true since Q_T is a positive random variable.

We now use Lemma C.2 by summing (C.11) from $t=2$ to T and (C.12) which leads to

$$\begin{aligned}
\mathbb{E}[Q_T\, \mathbb{1}_{H_{T-1}}] &\leq \mathbb{E}[Q_1\, \mathbb{1}_{H_0}] + \sum_{t=2}^{T} \gamma_t^2 \mathcal{M} - \sum_{t=2}^{T} \varepsilon\, \mathbb{P}(H_{t-2} \setminus H_{t-1}) \\
&\leq \mathbb{E}[Q_0\, \mathbb{1}_{H_{-1}}] + \gamma_1^2(2M^2 + 2\sigma^2 + 4r^2\sigma^2) + \sum_{t=2}^{T} \gamma_t^2 \mathcal{M} - \sum_{t=1}^{T} \varepsilon\, \mathbb{P}(H_{t-2} \setminus H_{t-1}) \\
&= \mathbb{E}[Q_0] + \gamma_1^2(2M^2 + 2\sigma^2 + 4r^2\sigma^2) + \sum_{t=2}^{T} \gamma_t^2 \mathcal{M} - \varepsilon\, \mathbb{P}({H_{T-1}}^c), \qquad \text{(C.22)}
\end{aligned}$$

where in the last line we use that $H_{-1} = \Omega$ and ${H_{T-1}}^c = H_{-1} \setminus H_{T-1} = \dot{\bigcup}_{1 \leq t \leq T}(H_{t-2} \setminus H_{t-1})$ (with $\dot{\bigcup}$ denoting the disjoint union) to get that $\mathbb{P}({H_{T-1}}^c) = \sum_{t=1}^{T} \mathbb{P}(H_{t-2} \setminus H_{t-1})$. Since we initialize with $X_{\frac{1}{2}} \in U$, we have

$$\mathbb{E}[Q_0] = 2\gamma_1^2\, \mathbb{E}[\|V_{\frac{1}{2}}\|^2] \leq 2\gamma_1^2(M^2 + \sigma^2). \qquad \text{(C.23)}$$

We set $\Gamma \coloneqq \sum_{t=1}^{\infty} \gamma_t^2 < \infty$. Combining (C.21), (C.22) and (C.23), we obtain

$$\begin{aligned}
\mathbb{P}({H_T}^c) &= \mathbb{P}(H_{T-1} \setminus H_T) + \mathbb{P}({H_{T-1}}^c) \\
&\leq \frac{1}{\varepsilon}\, \mathbb{E}[Q_T\, \mathbb{1}_{H_{T-1}}] + \mathbb{P}({H_{T-1}}^c) \\
&\leq \frac{1}{\varepsilon} \sum_{t=1}^{T} \gamma_t^2 \mathcal{M} - \mathbb{P}({H_{T-1}}^c) + \mathbb{P}({H_{T-1}}^c) \leq \frac{\Gamma \mathcal{M}}{\varepsilon}.
\end{aligned}$$

As Γ converges to 0 when $b \to \infty$, for any $\delta > 0$ one can choose b sufficiently large so that $\Gamma \leq \delta\varepsilon/\mathcal{M}$; we then have $\mathbb{P}({H_T}^c) \leq \delta$, or equivalently, $\mathbb{P}(H_T) \geq 1-\delta$ for all $T \geq -1$.

Since $H_{T-1} \subseteq E_T$ from Lemma C.1, we know that by choosing b sufficiently large, we have $\mathbb{P}(E_T) \geq \mathbb{P}(H_{T-1}) \geq 1-\delta$ for all $T \geq 0$. As $(E_T)_{T \geq 1}$ is a decreasing sequence of events and $E_\infty = \bigcap_{T \geq 0} E_T$, by continuity from above we have

$$\mathbb{P}(E_\infty) = \lim_{T \to \infty} \mathbb{P}(E_T) \geq 1 - \delta,$$

concluding the proof.

(b) Applying Lemma A.2 (b) gives

$$\begin{aligned}
\|X_{t+1} - x^\star\|^2 &\leq \|X_t - x^\star\|^2 - 2\gamma_t \langle V_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle \\
&\quad + \gamma_t^2 \|V_{t+\frac{1}{2}} - V_{t-\frac{1}{2}}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2 \\
&\leq \|X_t - x^\star\|^2 - 2\gamma_t \langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle \\
&\quad - 2\gamma_t \langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle \\
&\quad + 2\gamma_t^2(\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2) - \|X_{t+\frac{1}{2}} - X_t\|^2.
\end{aligned}$$

Furthermore, for any realization in E_t, $X_{t+\frac{1}{2}} \in U$ so that $\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle \geq \alpha \|X_{t+\frac{1}{2}} - x^\star\|^2$ and thus equation (C.6) holds, which allows us to write

$$\|X_{t+1} - x^\star\|^2\, \mathbb{1}_{E_t} \leq \|X_t - x^\star\|^2\, \mathbb{1}_{E_t} - 2\gamma_t \langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle\, \mathbb{1}_{E_t}$$

$$
\begin{aligned}
&- 2\gamma_t \langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle \mathbb{1}_{E_t} \\
&+ 2\gamma_t^2 (\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2) \mathbb{1}_{E_t} - \|X_{t+\frac{1}{2}} - X_t\|^2 \mathbb{1}_{E_t} \\
&\leq (1 - \alpha\gamma_t) \|X_t - x^\star\|^2 \mathbb{1}_{E_t} - 2\gamma_t \langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle \mathbb{1}_{E_t} \\
&+ 2\gamma_t^2 (\|V_{t+\frac{1}{2}}\|^2 + \|V_{t-\frac{1}{2}}\|^2) \mathbb{1}_{E_t} + (2\alpha\gamma_t - 1) \|X_{t+\frac{1}{2}} - X_t\|^2 \mathbb{1}_{E_t} . \quad \text{(C.24)}
\end{aligned}
$$

Similarly to (C.17) and (C.18), we have

$$
\begin{aligned}
&\mathbb{E}[\|V_{t+\frac{1}{2}}\|^2 \mathbb{1}_{E_t}] \leq M^2 + \sigma^2, \\
&\mathbb{E}[\|V_{t-\frac{1}{2}}\|^2 \mathbb{1}_{E_t}] \leq \mathbb{E}[\|V_{t-\frac{1}{2}}\|^2 \mathbb{1}_{E_{t-1}}] \leq M^2 + \sigma^2.
\end{aligned}
$$

We also recall that as $E_t \in \mathcal{F}_t$, it holds

$$
\mathbb{E}[\langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle \mathbb{1}_{E_t}] = \mathbb{E}[\mathbb{E}[\langle Z_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^\star \rangle \mid \mathcal{F}_t] \mathbb{1}_{E_t}] = 0.
$$

Taking expectation over (C.24) then leads to

$$
\begin{aligned}
\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mathbb{1}_{E_t}] \leq (1 - \alpha\gamma_t) \mathbb{E}[\|X_t - x^\star\|^2 \mathbb{1}_{E_t}] \\
+ 4\gamma_t^2 (M^2 + \sigma^2) + (2\alpha\gamma_t - 1) \mathbb{E}[\|X_{t+\frac{1}{2}} - X_t\|^2 \mathbb{1}_{E_t}].
\end{aligned}
$$

We can choose b sufficiently large so that $2\alpha\gamma_t - 1 \leq 0$ for all $t \geq 1$. Using $E_t \subseteq E_{t-1}$, we obtain

$$
\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mathbb{1}_{E_t}] \leq (1 - \alpha\gamma_t) \mathbb{E}[\|X_t - x^\star\|^2 \mathbb{1}_{E_{t-1}}] + 4\gamma_t^2 (M^2 + \sigma^2).
$$

By applying Lemma A.3 with $a_t \leftarrow \mathbb{E}[\|X_t - x^\star\|^2 \mathbb{1}_{E_{t-1}}]$, $q \leftarrow \alpha\gamma$, $q' \leftarrow 4\gamma^2(M^2 + \sigma^2)$, and $t_0 \leftarrow 1$, we get

$$
\mathbb{E}[\|X_t - x^\star\|^2 \mathbb{1}_{E_{t-1}}] \leq \frac{4\gamma^2(M^2 + \sigma^2)}{\alpha\gamma - 1} \frac{1}{t} + o\left(\frac{1}{t}\right).
$$

Finally,

$$
\begin{aligned}
\mathbb{E}[\|X_t - x^\star\|^2 \mid E_\infty] &= \frac{\mathbb{E}[\|X_t - x^\star\|^2 \mathbb{1}_{E_\infty}]}{\mathbb{P}(E_\infty)} \\
&\leq \frac{\mathbb{E}[\|X_t - x^\star\|^2 \mathbb{1}_{E_{t-1}}]}{1 - \delta} \\
&\leq \frac{4\gamma^2(M^2 + \sigma^2)}{(\alpha\gamma - 1)(1 - \delta)} \frac{1}{t} + o\left(\frac{1}{t}\right)
\end{aligned}
$$

and our proof is complete. □

Remark. We notice that to complete the above proof, we only require Eq. (2a) and Eq. (2b) to be held on the event $\{X_{t+\frac{1}{2}} \in U\}$. For example, in (C.16) we want $\mathbb{E}[\mathbb{E}[\|Z_{t+\frac{1}{2}}\|^2 \mid \mathcal{F}_t] \mathbb{1}_{H_{t-1}}] \leq \sigma^2$ which is true if Eq. (2b) holds on $\{X_{t+\frac{1}{2}} \in U\}$ since $H_{t-1} \subseteq E_t \subseteq \{X_{t+\frac{1}{2}} \in U\}$. This assumption is much weaker and more sensible. It in particular shows that to obtain local guarantee we indeed only need the noise to be bounded locally.