Sifting through the Noise: Universal First-Order Methods for Stochastic Variational Inequalities

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

Thomas Pethick **Ali Kavis**
École Polytechnique Fédérale de Lausanne (EPFL)
thomas.pethick@epfl.ch ali.kavis@epfl.ch

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

Volkan Cevher
École Polytechnique Fédérale de Lausanne (EPFL)
volkan.cevher@epfl.ch

Abstract

We examine a flexible algorithmic framework for solving monotone variational inequalities in the presence of randomness and uncertainty. The proposed template encompasses a wide range of popular first-order methods, including dual averaging, dual extrapolation and optimistic gradient algorithms – both adaptive and non-adaptive. Our first result is that the algorithm achieves the optimal rates of convergence for cocoercive problems when the profile of the randomness is known to the optimizer: $\mathcal{O}(1/\sqrt{T})$ for absolute noise profiles, and $\mathcal{O}(1/T)$ for relative ones. Subsequently, we drop all prior knowledge requirements (the absolute/relative variance of the randomness affecting the problem, the operator's cocoercivity constant, etc.), and we analyze an adaptive instance of the method that gracefully interpolates between the above rates – i.e., it achieves $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T)$ in the absolute and relative cases, respectively. To our knowledge, this is the first universality result of its kind in the literature and, somewhat surprisingly, it shows that an extra-gradient proxy step is not required to achieve optimal rates.

1 Introduction

This paper focuses on solving variational inequality problems of the form

$$\text{Find } x^* \in \mathbb{R}^d \text{ such that } \langle A(x^*), x - x^* \rangle \geq 0 \text{ for all } x \in \mathbb{R}^d, \tag{VI}$$

where $A\colon \mathbb{R}^d \to \mathbb{R}^d$ is a *monotone cocoercive* operator, i.e.,

$$\langle A(y) - A(x), y - x \rangle \geq \beta \|A(y) - A(x)\|^2 \quad \text{for some } \beta > 0 \text{ and all } x, y \in \mathbb{R}^d. \tag{CC}$$

The study of monotone VI problems is a classical topic in optimization that provides a powerful and elegant unifying framework for a broad spectrum of "convex-structured" problems – including convex minimization, saddle-point problems, and games [4, 12].

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

The VI problems in general have recently attracted considerable attention in the fields of machine learning (ML) and data sciences because of their potential applications to generative adversarial networks [15], multi-agent and robust reinforcement learning [34], auction theory [41], and many other areas of interest where the minimization of a single empirical loss function does not suffice.

Stochastic first-order methods are the de facto standard in this setting: these methods can be run with computationally cheap updates that only require noisy access to A, so they are ideal for problems with very high dimensionality and moderate-to-low precision needs as is typically the case in ML.

When A is monotone cocoercive as above, the min-max optimal convergence rate for solving (VI) is $\mathcal{O}(1/T)$ after T oracle calls, and it is achieved by the extra-gradient / mirror-prox algorithm [23, 27] with Polyak-Rupert averaging [36]. However, this method requires access to a *perfect* oracle; if the method is run with a stochastic first-order oracle, its convergence rate deteriorates to $\mathcal{O}(1/\sqrt{T})$ [21], and this rate is, in general, not improvable [29] without additional assumptions.

The $\mathcal{O}(1/\sqrt{T})$-rate can be improved when the underlying operator is *strongly monotone* – i.e., the RHS of (CC) is replaced by $\alpha\|y-x\|^2$ for some $\alpha > 0$. In this case, we can obtain a fast $\mathcal{O}(1/T)$ rate with a rapidly decreasing step-size [16]; however, this acceleration requires knowledge of the strong monotonicity modulus, and there is no known way to adapt to it. In particular, if a stochastic first-order method that has been fine-tuned for strongly monotone operators is run on a merely monotone/cocoercive problem, its rate of convergence suffers a catastrophic drop to $\mathcal{O}(1/\log T)$.

These considerations naturally lead two key research questions: First, *are there any conditions for the method's oracle that would close the stochastic-deterministic convergence gap outlined above?* Second, *is it possible to design a class of methods that would adapt to the quality of the oracle, and that achieve order-optimal rates without prior knowledge of the problem's parameters* (the operator's cocoercivity modulus, the variance of the oracle, etc.)?

Our contributions in the context of related work. We provide a range of positive answers to these questions, both in terms of oracle requirements, as well as methods that are able to gracefully interpolate between an $\mathcal{O}(1/T)$ and an $\mathcal{O}(1/\sqrt{T})$ rate, depending on the setting at hand.

With regard to the first question, our point of departure is the classical "relative noise" framework of Polyak [35], in which the variance of the oracle is upper bounded by the square norm of the operator at the queried point. This noise model is particularly relevant in applications to control theory and signal processing, when the operator is calculated based on actual, physical measurements that are only accurate up to a percentage of their true value. In recent applications to ML, this noise model has also been studied in the context of overparametrization [32], representation learning [46], and multi-agent learning [25]. This model has also been studied under the umbrella of multiplicative noise [19] or growth conditions [5, 39, 42, 44], and it is known to improve the convergence rate of stochastic gradient algorithms with non-adaptive step-sizes, even in non-smooth problems [13].

With regard to the second question, we introduce a flexible first-order algorithmic template that includes as special cases the dual averaging [31], dual extrapolation [30] and optimistic gradient methods [37, 38], and accounts for both adaptive and non-adaptive variants thereof. Our contributions can then be summarized as follows:

1. For oracles with bounded variance, we show that the proposed methods achieve an $\mathcal{O}(1/\sqrt{T})$ if run with a non-adaptive, decreasing step-size.
2. In the relative noise model, this rate improves to $\mathcal{O}(1/T)$, and it is achieved with a *constant* step-size that does need to be tuned as a function of T.
3. Finally, we provide an adaptive step-size rule that allows the method to achieve a fast, $\mathcal{O}(1/T)$ rate under relative noise, and an order-optimal $\mathcal{O}(1/\sqrt{T})$ in the absolute noise case.

Importantly, our work shows that an extra-gradient mechanism is *not* required to obtain a fast $\mathcal{O}(1/T)$ rate, as this can be achieved by vanilla dual-averaging methods with a constant step-size. This is an elegant consequence of the interplay between cocoercivity and the relative noise model; to the best of our knowledge, the only other work considering these models in tandem is the very recent paper [25].

Our work closes several open threads in [25], which requires a *vanishing* relative noise level to obtain faster convergence in models with relative noise. A summary of our results in the context of related work can be found in Table 1. Appendix A also elaborates on the related work in greater detail.

	V_t	Lipschitz		Cocoercive + rel. noise	
		Ergodic	Last Iterate	Ergodic	Last Iterate
Adapt. dual averaging	0	$1/\sqrt{T}$ [10]	Unknown	$1/T$	Asym.
Adapt. dual extrapolation	AX_t + rel.noise	$1/\sqrt{T}$ [38]	Unknown	$1/T$	Asym.
Adapt. optimistic gradient	$AX_{t-1/2}$ + rel.noise	$1/\sqrt{T}$ [11]	Unknown	$1/T$	Asym.

Table 1: The best known convergence rates in stochastic monotone VIs with our contributions highlighted in gray. *Adaptive* refers to our particular adaptive step-size choice in (Adapt). We obtain various schemes with particular choices of V_t. For the nomenclature, please refer to Section 3.2.

2 Problem setup and preliminaries

Throughout the sequel, we will focus on solving the variational inequality problem (VI). We briefly mention some examples below, and we defer to [12, 40] for a panoramic survey of the field.

Example 1. Convex Minimization: If $A = \nabla f$ for some convex function f, then the solutions of the (VI) are simultaneously the minimizers of f and vice versa.

Example 2. Min-Max: If $A = (\nabla_{x_1} L, -\nabla_{x_2} L)$ for some convex-concave function $L(x_1, x_2)$, then the solutions of (VI) coincide with the (global) saddle points of L. More precisely, $x^* = (x_1^*, x_2^*)$ is a solution of (VI) if and only if it holds that

$$L(x_1^*, x_2) \leq L(x_1^*, x_2^*) \leq L(x_1, x_2^*) \quad \text{for all} \quad x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2. \tag{SP}$$

In this case, (VI) is sometimes referred to as the vector field formulation of (SP).

Example 3. Convex Games: Beyond min-max problems, a continuous game in normal form can be stated formally as follows: Consider a finite set of players $\mathcal{N} = \{1, \dots, N\}$ which each player choosing actions $x_i \in \mathbb{R}^{d_i}$. In particular, each player seeks to minimize an individual loss $l_i(x_i, x_{-i})$ which is determined by the player's action along with the actions of the adversaries.

In terms of regularity conditions for the associated operator, we assume that A is β- cocoercive (CC); for a panoramic overview of cocoercive operators we refer the reader to [4].

Some further comments for the cocoercivity condition are in order. First, one may easily observe that if A is β-cocoercive, then A is also $1/\beta$-Lipschitz. However, the converse does not hold for the general setting of operators; however for smooth convex this is actually the case [3]. Moreover, cocoercivity, although it yields monotonicity, does imply in general strictly monotone a condition which usually applied in order to ensure the existence and uniqueness of a (VI) solution. Therefore, to avoid pathologies, we make the following assumption for our setting:

Assumption 1. The solution set $\mathcal{X}^* = \{x^* \in \mathbb{R}^d : x^* \text{ is a solution of (VI)}\}$ is non-empty.

With the above setup in hand, a widely used performance measure in order to evaluate a candidate solution of (VI) is that of the so-called *restricted gap function*:

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

where the "test domain" $\mathcal{C}$ is a non-empty compact subset of $\mathbb{R}^d$. The motivation for the choice of (Gap) is that it characterizes the solutions of the (VI) via its zeros. Formally, we have the following:

Proposition 1. *Let $\mathcal{C}$ be a non-empty convex subset of $\mathbb{R}^d$. Then, the following holds*

1. $\mathrm{Gap}_{\mathcal{C}}(\hat{x}) \geq 0$, *whenever* $\hat{x} \in \mathcal{C}$
2. *If* $\mathrm{Gap}_{\mathcal{C}}(\hat{x}) = 0$ *and $\mathcal{C}$ contains a neighbourhood of $\hat{x}$, then $\hat{x}$ is a solution of* (VI)

Proposition 1 is in turn a generalization of an earlier characterization by Nesterov in [30]. Moreover, it provides a formal justification for the use of $\mathrm{Gap}_{\mathcal{C}}(\hat{x})$ as a merit function for (VI). To streamline our presentation we defer the details to the paper's supplement.

3 The method

3.1. Oracle structure and different profiles of randomness. From an algorithmic point of view, in order to solve (VI) we will use iterative methods that require access to a stochastic first order oracle [29]. Formally, this is a black-box feedback mechanism which when called at x returns a random

dual vector $g(x;\omega)$ with ω drawn from some (complete) probability space $(\Omega, \mathcal{F}, \mathbb{P})$: In practice, the oracle will be repeatedly called at a possibly random sequence of points generated by the algorithm at play. Therefore, once the iterate of the method is generated at each round, the oracle i.i.d. sample $\omega \in \Omega$ and returns a dual vector:

$$g(x;\omega) = A(x) + U(x;\omega) \tag{1}$$

with $U(x;\omega)$ denoting the "measurement error". Having all this at hand, we make the following blanket statistical assumptions:

Assumption 2. 1. (Almost sure boundedness): There exists some strictly positive numbers $M > 0$ such that:

$$\|g(x;\omega)\|_* \leq M \quad \text{almost surely} \tag{2}$$

2. (Unbiasedness): $\mathbb{E}\left[g(x;\omega)\right] = A(x)$

3. (Bounded Variance): $\mathbb{E}\left[\|U(x;\omega)\|_*^2\right] \leq \sigma^2$

Such type of conditions for the oracle are typical especially for adaptive methods; see for example [2, 22, 24] and we shall denote them as *absolutely random*. On the other hand, we shall also consider the so-called *relatively random* [35]. More precisely, we have the following:

Assumption 3. 1. (Almost sure boundedness): There exists some strictly positive numbers $M > 0$ such that: $\|g(x;\omega)\|_* \leq M$ almost surely

2. (Unbiasedness): $\mathbb{E}\left[g(x;\omega)\right] = A(x)$

3. (Relative Variance): There exists some positive $c > 0$ such that:

$$\mathbb{E}\left[\|U(x;\omega)\|_*^2\right] \leq c\|A(x)\|_*^2 \tag{3}$$

As a prelude of what is going to follow some comments are in order. It is well known that Assumption 2 is a typical assumption in order to get the typical stochastic $\mathcal{O}(1/\sqrt{T})$ for various optimization scenarios (see for example [21, 28] and references therein). That said, Assumption 3 will prove itself as the crucial statistical condition that will allow us to recover the well known order-optimal bound $\mathcal{O}(1/T)$ for deterministic settings.

3.2. The methods. We now present the generalized extra-gradient (GEG) family of algorithms. More precisely, given two sequences of dual vectors V_t and $V_{t+1/2}$ (GEG) is given by the following recursive formula:

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t V_t \\ Y_{t+1} &= Y_t - V_{t+1/2} \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{GEG}$$

Heuristically, the machinery behind (GEG) suggests to first generate a leading state $X_{t+1/2}$ via a gradient descent method towards the direction of the dual sequence of V_t, then aggregate the feedback by incorporating the second dual sequence $V_{t+1/2}$ and finally update the method by applying a dual averaging step [31, 43]. This idea is fairly standard in the literature of extra-gradient methods [23, 27, 30]. However, up to this point, we have not assumed anything particular for the sequences of V_t and $V_{t+1/2}$, except that they are dual vectors (but not necessarily queries of a stochastic oracle of (1)). This generic choice is the building block that will allow us to include various popular algorithmic schemes and provide a unified framework fo their analysis.

To begin with, we provide the following examples that illustrate the fact that Dual Averaging, Dual Extrapolation and Optimistic Dual Averaging all can be written in the form of (GEG) under different choices of V_t and $V_{t+1/2}$.

Example 4. Stochastic Dual Averaging [31]: Consider the case where $V_t \equiv 0$ and $V_{t+1/2} \equiv g_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$. Then, this yields that $X_{t+1/2} = X_t$ and hence $g_{t+1/2} = g_t = V_{t+1/2}$. Therefore, the (GEG) scheme reduces to the dual averaging scheme:

$$\begin{aligned} Y_{t+1} &= Y_t - g_t \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{DA}$$

Example 5. Stochastic Dual Extrapolation [30]: Consider the case now where $V_t \equiv g_t = A(X_t) + U_t$ and $V_{t+1/2} \equiv g_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ are a "noisy" oracle feedbacks at X_t and $X_{t+1/2}$ respectively, then directly (GEG) yields Nesterov's dual extrapolation method:

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t g_t \\ Y_{t+1} &= Y_t - g_{t+1/2} \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{DE}$$

Example 6. Stochastic Optimistic Dual Averaging [37, 38]: Consider for the case where $V_t \equiv g_{t-1/2} = A(X_{t-1/2}) + U_{t-1/2}$ and $V_{t+1/2} \equiv g_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ are the noisy oracle feedback at $X_{t-1/2}$ and $X_{t+1/2}$ respectively. We then get the optimistic dual averaging method:

$$\begin{aligned} X_{t+1/2} &= X_t - \gamma_t g_{t-1/2} \\ Y_{t+1} &= Y_t - g_{t+1/2} \\ X_{t+1} &= \gamma_{t+1} Y_{t+1} \end{aligned} \tag{OptDA}$$

The next crucial step is to provide the key ingredient that will allow us to unify the approach for all algorithms belonging to the family (GEG). This is done by a shared "energy" inequality satisfied by all (GEG)-type schemes. Formally, this is described by the following proposition:

Proposition 2. *Assume that $X_t, X_{t+1/2}$ are the iterates of* (GEG) *run with a non-negative, non-increasing step-size γ_t. Then, for all $x \in \mathbb{R}^d$ the following inequality holds:*

$$\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x \right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \left\|V_{t+1/2} - V_t\right\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t} \left\|X_{t+1/2} - X_t\right\|^2 \tag{4}$$

Proving Proposition 2 requires tiresome computations, so we defer it to the paper's supplement. Instead, we conclude this section by illustrating the various method-specific template inequalities:

1. (Stochastic Dual Averaging): For $V_{t+1/2} = g_{t+1/2}$ and $V_t = 0$, then (4) becomes:

$$\sum_{t=1}^{T} \langle g_t, X_t - x \rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \|V_t\|_*^2 \tag{5}$$

2. (Stochastic Dual Extrapolation): For $V_t = g_t$ for all $t = 1, 1/2, \ldots$ then (4) becomes:

$$\sum_{t=1}^{T} \left\langle g_{t+1/2}, X_{t+1/2} - x \right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \left\|g_{t+1/2} - g_t\right\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t} \left\|X_{t+1/2} - X_t\right\|^2 \tag{6}$$

3. (Stochastic Optimistic Dual Averaging): For $V_t = g_{t-1/2}$ and $V_{t+1/2} = g_{t+1/2}$ then (4) becomes:

$$\sum_{t=1}^{T} \left\langle g_{t+1/2}, X_{t+1/2} - x \right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t \left\|g_{t+1/2} - g_{t-1/2}\right\|_*^2 - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t} \left\|X_{t+1/2} - X_t\right\|^2 \tag{7}$$

4 Non-adaptive Analysis

In this section, we derive a series of tight convergence rates for the family (GEG) under both oracle/noise profiles but with a *non-adaptive* step-size sequences. Due to space constraints, we defer the full analysis to the supplement; however, we provide here a proof sketch of our main results via an appropriate "energy inequality" in Proposition 2.

4.1. Absolute random noise. In the context of monotone VIs, assumptions induced by the random oracle model are common and well-understood. Indeed, for the general case of bounded variance, i.e., $\mathbb{E}\left[U_{t+1/2} | \mathcal{F}_{t+1/2}\right] \leq \sigma$, extra-gradient/mirror-prox is known to converge at a rate $\mathcal{O}(1/\sqrt{T})$ [20], with a decreasing step-size of order $\mathcal{O}(1/\sqrt{t})$.

For completeness, we analyze (GEG) under a random oracle profile, i.e., for $V_{t+1/2} = g_{t+1/2} \equiv g(X_{t+1/2}; \omega_{t+1/2})$ satisfying Assumption 2 and V_t being an almost surely bounded sequence of dual vectors. To that end, we employ a decreasing step-size choice, which is summarized in the next theorem.

Theorem 1. *Let $X_t, X_{t+1/2}$ be generated by (GEG) with a decreasing step-size $\gamma_t = \mathcal{O}(1/\sqrt{t})$. Then, for every compact neighborhood $\mathcal{C} \subset \mathbb{R}_d$ of x^*, with $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_{t+1/2}$, it holds that:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\bar{X}_T\right)\right] = \mathcal{O}(1/\sqrt{T}).$$

The arguments for the proof of Theorem 1 are standard and we defer them to the appendix due to space constraints. Thanks to this result, we can now derive the respective method specific rates as special instances. More precisely, we have the following proposition:

Proposition 3. *Under Assumption 2 the iterates of (DA), (DE), (OptDA) enjoy the following rate:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\bar{X}_T\right)\right] = \mathcal{O}(1/\sqrt{T}) \tag{8}$$

4.2. Relative random noise. We now turn our attention to the relative random oracle framework, i.e. $V_{t+1/2} = g_{t+1/2}$ satisfying Assumption 2 along with:

$$\mathbb{E}\left[\|V_t\|_*^2 | \mathcal{F}_t\right] \leq c\|A(X_t)\|_*^2 \quad \text{for all} \quad t = 1, 1/2, \dots \tag{9}$$

In particular, with a carefully chosen constant step-size, under the additional assumption of relative variance, it is possible to achieve an accelerated rate of $\mathcal{O}(1/T)$. One needs to depart from the standard approach to fully exploit the problem setting, i.e., cocoercivity and relative variance. Essentially, it amounts to ensuring that $\sum_{t=1}^{T} \|A_t\|_*^2$ and $\sum_{t=1}^{T} \|A_{t+1/2}\|_*^2$ are summable. We present our result under the respective setting with a proof sketch that highlights its main ingredients.

Theorem 2. *Let $X_t, X_{t+1/2}$ be generated by (GEG) with a constant step-size that satisfies*

$$\min\left\{(2L)^{-1}, (4L^2\gamma)^{-1}\right\} - 2\gamma c > 0. \tag{10}$$

Then, for every compact neighbourhood $\mathcal{C} \subset \mathbb{R}_d$ of x^*, with $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_{t+1/2}$, we have:

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\bar{X}_T\right)\right] = \mathbb{E}\left[\sup_{X \in \mathcal{C}} \left\langle A(X), \bar{X}_T - X\right\rangle\right] = \mathcal{O}(1/T)$$

Proof. With a constant step-size, Proposition 2 implies

$$\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - X\right\rangle = \frac{\|X\|^2}{2\gamma} + \frac{\gamma}{2}\sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma}\sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2$$

We show that using smoothness and cocoercivity of the operator, along with the relative noise condition,

$$\left(\min\left\{(2L)^{-1}, (4L^2\gamma)^{-1}\right\} - 2\gamma c\right)\sum_{t=1}^{T}\left(\mathbb{E}\left[\|A(X_t)\|^2\right] + \mathbb{E}\left[\|A(X_{t+1/2})\|^2\right]\right) \leq \frac{\mathbb{E}\left[\|X\|^2\right]}{\gamma}$$

If constant step-size γ satisfies Eq. (10), then there exists some strictly positive real number β, such that $\mathbb{E}\left[\sum_{t=1}^{T}\left(\|A(X_t)\|^2 + \|A(X_{t+1/2})\|^2\right)\right] \leq \mathbb{E}\left[\|X\|^2/\beta\gamma\right] < +\infty$, which concludes that both $\sum_{t=1}^{T} \|A_t\|_*^2$ and $\sum_{t=1}^{T} \|A_{t+1/2}\|_*^2$ are summable. Using the same arguments as in the proof of Theorem 1, we obtain an upper bound for the gap,

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_{T+1/2})\right] \leq \frac{\frac{D^2}{2\gamma} + 2\gamma c\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right] + \sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|V_{t+1/2}\right\|_*^2\right]}}{T}.$$

By relative variance and summability of operators,

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_{T+1/2})\right] = \mathcal{O}(1/T)$$

□

Similar to the setting of absolutely random noise, Theorem 2 implies algorithm-specific convergence bounds, which are presented below:

Proposition 4. *Under Assumption 3 the iterates of* (DA)*,* (DE)*,* (OptDA) *enjoy the following rate:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\bar{X}_T\right)\right] = \mathcal{O}(1/T) \tag{11}$$

An extra appealing feature of the above is that we are able to derive an asymptotic last iterate trajectory result, i.e., the asymptotic convergence of the iterates themselves before any averaging occurs, almost surely. More precisely, we have the following proposition:

Proposition 5. *Under Assumption 3 the iterates of* (DA)*,* (DE)*,* (OptDA) *converge to a* (VI) *solution* x^*.

The proof Proposition 5 relies on the fact that the distance of the iterates towards any solution of (VI) is decreasing almost surely along with the fact that the summability of $\|A(X_t)\|_*^2$ guarantees that every limit point of the iterate is also a solution of (VI). Due to space constraints we defer the detailed proof to the supplement.

5 Adaptive Analysis

By the results of Section 4, one may easily observe the interplay between the $\mathcal{O}(1/\sqrt{T})$ to $\mathcal{O}(1/T)$ rate interpolations under different noise profiles and step-sizes policies. Therefore a natural question that arises from this context is the following:

Could we apply one single step-size policy that is able to optimally adjust the (GEG) *performance without any prior knowledge over the noise profile structure?*

In what follows this desired property is accomplished if the (GEG) algorithms are run with the following adaptive step-size:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|V_j - V_{j+1/2}\|_*^2}} \tag{Adapt}$$

The step-size (Adapt) is inspired by [38]; however in our analysis we provide a generalized point of view since we do not assume that V_t necessarily is the oracle query at the respective points as in [38]. This allows us to include in the (Adapt) formulation all the adaptive sttep-sizes typically used for the archetypical schemes introduced in Section 3 . More precisely, we have:

1. (Adaptive Stochastic Dual Averaging): For $V_t \equiv 0$ (Adapt) becomes the standard AdaNorm stepsize, studied in various works (see. for example, [10, 26]:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|g_j\|_*^2}} \tag{12}$$

2. (Adaptive Stochastic Dual Extrapolation): For $V_t = g_{t+1/2}$ (Adapt) becomes

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|g_j - g_{j+1/2}\|_*^2}} \tag{13}$$

   used in [1, 38].

3. (Adaptive Stochastic Optimistic Dual Averaging):. For $V_t = g_{t-1/2}$ (Adapt) becomes the step-size used in [17]:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|g_{j+1/2} - g_{j-1/2}\|_*^2}} \tag{14}$$

Section 4, heuristically suggests that the success of γ_t should hinge on a simultaneous performance as $1/\sqrt{t}$ for the absolute random oracle feedback and as a constant one. whenever the relative random feedback kicks in. This important interpolation feature is what will show in thhe sequel.

5.1. Absolutely random noise. We first will treat the absolutely random noise. In particular, we have the following result

Theorem 3. *Assume that $X_t, X_{t+1/2}$ are the* (GEG) *iterates run with the step-size* (Adapt)*. Then, for every compact neighbourhood $\mathcal{C} \subset \mathbb{R}^d$ of a* (VI) *solution x^*, the following inequality holds:*

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{15}$$

with $\overline{X}_T = 1/T \sum_{t=1}^{T} X_{t+1/2}$

As we argued above, the result of Theorem 3 is heuristically justified by the fact that the almost sure boundedness conditions for the sequences:

$$\|V_t\|_* \leq M \text{ almost surely for all } t = 1, 1/2, \ldots \tag{16}$$

yields that $\gamma_t = \Omega(1/\sqrt{t})$. Moreover, as desired, Theorem 3 yields the method specific rates of convergence:

Proposition 6. *Under Assumption 2 the iterates of* (DA)*,* (DE)*,* (OptDA) *enjoy the following:*

$$\mathbb{E}\left[Gap_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{17}$$

5.2. Relative random noise. Under the relative random noise conditions, we can obtained the following improved rate $\mathcal{O}(1/T)$ instead of the typical $\mathcal{O}(1/\sqrt{T})$. Formally, we have the following theorem.

Theorem 4. *Assume that $X_t, X_{t+1/2}$ are the* (GEG) *iterates run with the step-size* (Adapt)*. Then, for every compact neighbourhood $\mathcal{C} \subset \mathbb{R}^d$ of a* (VI) *solution x^*, the following inequality holds:*

$$\mathbb{E}\left[Gap_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/T) \tag{18}$$

with $\overline{X}_T = 1/T \sum_{t=1}^{T} X_{t+1/2}$

The crucial ingredient for the proof of Theorem 4 consists of the fact that the adaptive step size stabilizes to a positive constant $\gamma_\infty > 0$. In order to obtain this, by applying the template inequality obtained in Proposition 2 , we show that:

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \leq \left(8c \max\left\{L, 2L^2\right\}\left(\frac{\|x^* - x_1\|^2}{2} + 2G^2 + 1\right) + 1\right)\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{19}$$

Moreover, due to the definition of (Adapt) and Jensen's inequality we have:

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] = \mathbb{E}\left[\sqrt{1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|^2}\right] \leq \sqrt{\mathbb{E}\left[1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|^2\right]} = \sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]} \tag{20}$$

Therefore, after combining (19) and (20) we get that $\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] < +\infty$. This directly implies (due to the Monotone Convergence Theorem) that:

$$\frac{1}{\gamma_{T+1}^2} = 1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2 < +\infty \text{ almost surely} \tag{21}$$

which in turn yields that $\sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2$ is summable almost surely. Therefore due to the definition of γ_t we have almost surely the following:

$$\gamma_{T+1} = \frac{1}{\sqrt{1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2}} \to \frac{1}{\sqrt{1 + \sum_{t=1}^{+\infty}\|V_t - V_{t+1/2}\|_*^2}} = \gamma_\infty > 0 \tag{22}$$

Finally, we conclude by providing the following the respective method specific result. Formally, we have:

Proposition 7. *Under Assumption 3 the iterates of* (DA)*,* (DE)*,* (OptDA) *enjoy the following:*

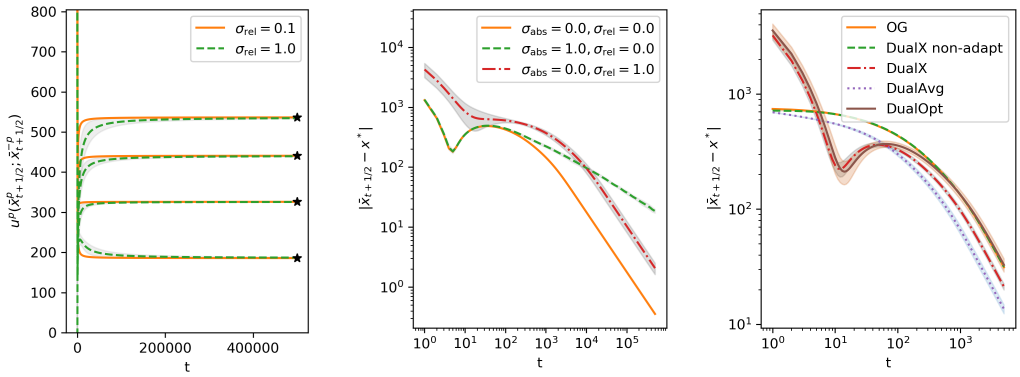


Figure 1: (left) Player utility using adaptive DualX for various relative noise levels. Even at relative high levels of noise do we converge to the optimal depicted with ($\star$). (center) Average iterate for deterministic, absolute noise and relative noise using adaptive DualX. We observe the $\mathcal{O}(1/\sqrt{T})$ rate under absolute noise while $\mathcal{O}(1/T)$ is achieved both in the noiseless setting and under relative noise. In addition, as expected, the last iterate only converges under the deterministic and relative noise oracle (see Fig. H.2). (right) Average iterate comparing various methods for $\sigma_{\text{rel}} = 0.1$. All methods shares convergence rate with adaptive methods being slightly faster possibly because of difficulty of step-size tuning for non-adaptive methods. Error bars indicate one standard deviation computed using 10 independent executions.

1. *The convergence rate in terms of the restricted gap function for the time-average:*

$$\mathbb{E}\left[Gap_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/T) \tag{23}$$

2. *Their last iterate trajectory converges to a* (VI) *solution* x^* *almost surely.*

The last iterate convergence result of Proposition 7 refers to the asymptotic convergence of the actual sequences of the methods-before any averaging takes place- and it hinges on the fact that the (random) sequences $\|V_t - V_{t+1/2}\|_*^2$ and $\|A(X_t)\|_*^2$ are summable with probability 1. Having established this, we show that X_t satisfies is a (stochastic) quasi-Fejér sequence [7] (with respect to the solution set $\mathcal{X}^*$) along with the fact that every limit point of X_t belongs to $\mathcal{X}^*$. These two building blocks are sufficient in order to derive the almost sure convergence of the iterate's trajectory.

6 Numerical experiments

In this section we validate and explore the consequences of the theoretical results. We adopt the experimental setting considered in [14] which is a particular instance of the Kelly auction with $N = 4$. In its generality in a single resource Kelly auction, there are N players sharing a total amount of $Q \in \mathbb{R}_{>0}$ resources. At every round, each bidder, p, submits a bid $x^p \in \mathbb{R}_{\geq 0}$ and receives proportional resources, $\rho^p = \frac{Qx^p}{Z + \sum_p x^p}$, where Z is the auction entry price. The payoff for player p is then given as $u^p(x^p; x^{-p}) = G^p \rho^p - x^p$, where G^p is the marginal gain in utility for player p. One can easily verify that the vectorfield associated with the payoff functions is cocoercive. In addition, the assumption of relative noise can be justified since each player can be seen as performing a measurement when querying the payoff. In such settings, it is common to assume that the error is proportional to the measured quantity and this uncertainty propagates to the gradient information in the form of relative noise. Since players act without communication in this example, it is particularly important that our results extends to single-call extragradient variants (see for instance [38] for elaboration). However, note that our proposed adaptive step-size (Adapt) still relies on global information of all players so our non-adaptive results for known problem constants is also important for this example.

In order to simulate the presence of relative noise we add a term proportional to the norm of the operator. In our notation we can thus capture both relative noise and absolute noise through the error term U_t in the following way,

$$U_t = \epsilon_{\text{rel}} \|A(X_t)\| + \epsilon_{\text{abs}}, \tag{24}$$

where $\epsilon_{\text{rel}} \sim \mathcal{N}(0, \sigma_{\text{rel}}^2)$ and $\epsilon_{\text{abs}} \sim \mathcal{N}(0, \sigma_{\text{abs}}^2)$. To validate the convergence rate we compute the optimal strategy in the deterministic setting (i.e. $\sigma_{\text{rel}} = \sigma_{\text{abs}} = 0$) using Mathematica.

In Fig. 1 we illustrate the behavior of the different instantiations of our algorithmic template under different choices of σ_{rel} and σ_{abs}. To denote (DA), (DE) and (OptDA) we use DualAvg, DualX and DualOpt respectively. In addition we include optimistic gradient (OG) from [9] for comparison. For higher dimensional experiments see Appendix H.1. In Appendix H.2 we additionally apply our adaptive method to the non-convex problem of learning a covariance matrix introduced in [9].

7 Concluding remarks

In this paper we provide rate interpolation guarantees for different noise profiles; namely that of absolute and relative random noise. That being said our analysis crucially depends on the cocoercivity of the associated operator that defines the respective (VI). It thus remains open whether it is possible to achieve the same $\mathcal{O}(1/T)$ rate for monotone (VI) by only assuming Lipschitz continuity of the said operator and relative noise. Moreover, an additional interesting direction for future research is investigate the impact of relative noise for adaptive accelerated methods and whether it is possible to recover the iconic $\mathcal{O}(1/T^2)$ rate. We delegate these questions for future works.

Acknowledgments and Disclosure of Funding

This research was partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET) and the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the grant ALIAS (ANR-19-CE48-0018-01). On the EPFL side, this project has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data), the Hasler Foundation Program: Cyber Human Systems (project number 16066), and the Swiss National Science Foundation (SNSF) under grant number $200021178865/1$. This project was sponsored by the Department of the Navy, Office of Naval Research (ONR) under grant number N62909-17-1-2111, and the Army Research Office under Grant Number W911NF-19-1-0404.

References

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

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

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

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

[5] Volkan Cevher and Bang Cong Vu. On the linear convergence of the stochastic gradient method with constant step-size. *arXiv:1712.01906 [math]*, June 2018.

[6] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing Noise in GAN Training with Variance Reduced Extragradient. *arXiv:1904.08598 [cs, math, stat]*, June 2020.

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

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

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

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[11] Alina Ene and Huy L Nguyen. Adaptive and universal single-gradient algorithms for variational inequalities. *arXiv preprint arXiv:2010.07799*, 2020.

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

[13] Huang Fang, Zhenan Fan, and Michael P Friedlander. Fast convergence of stochastic subgradient method under interpolation. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[14] Bolin Gao and Lacra Pavel. Discounted mirror descent dynamics in concave games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5942–5947. IEEE, 2019.

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

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

[17] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In *COLT' 21,Proceedings of 34th Annual Conference on Learning Theory*, 2021.

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

[19] Alfredo N Iusem, Alejandro Jofré, Roberto I Oliveira, and Philip Thompson. Variance-based extragradient methods with line search for stochastic variational inequalities. *SIAM Journal on Optimization*, 29(1): 175–206, 2019.

[20] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17 – 58, 2011. doi: 10.1214/10-SSY011. URL https://doi.org/10.1214/10-SSY011.

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

[22] Ali Kavis, Kfir Yehuda Levy, Francis Bach, and Volkan Cevher. UnixGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

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

[24] Kfir Yehuda Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[25] Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael I. Jordan. Finite-time last-iterate convergence for multi-agent learning in games. In *ICML '20: Proceedings of the 37th International Conference on Machine Learning*, 2020.

[26] H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *COLT '10: Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

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

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

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

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

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

[32] Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1 (1):84–105, 2020.

[33] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/1aa48fc4880bb0c9b8a3bf979d3b917e-Paper.pdf.

[34] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017.

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

[36] Boris Teodorovich Polyak and Anatoli Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 1992.

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

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

[39] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

[40] Gesualdo Scutari, Francisco Facchinei, Daniel Pérez Palomar, and Jong-Shi Pang. Convex optimization, game theory, and variational inequality theory in multiuser communication systems. *IEEE Signal Process. Mag.*, 27(3):35–49, May 2010.

[41] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *NIPS '15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, pages 2989–2997, 2015.

[42] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron. *arXiv:1810.07288 [cs, stat]*, April 2019.

[43] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, October 2010.

[44] Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear convergence of adaptive stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1475–1485. PMLR, 2020.

[45] Farzad Yousefian, Angelia Nedic, and Uday V. Shanbhag. Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. *arXiv:1403.5591 [math]*, March 2014.

[46] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Further related work

Due to space limitations for the main paper, we provide in this section a more detailed panorama of the related work.

Adaptivity Adaptive schemes that achieve optimal rates even without knowing the noise constant have been considered before in the $\min\text{-}\max$ optimization setting [2]. However, [2] focuses on the general case where only $\mathcal{O}(1/\sqrt{T})$ is possible. It is also worth pointing out that their step-size relies on the gradient mapping since they consider constrained $\min\text{-}\max$ problems, while ours is based on the operator difference since we consider unconstrained VI problems.

In this sense, [2] is closer to the scheme in [1], where they focus on adapting to non-smooth/smooth problems with unbounded domains in the *deterministic* setting. For the stochastic setting, there exists results for a single-call method using the same adaptive step-size as ours [11]. This work allows us to recover the $\mathcal{O}(T^{-1/2})$ convergence in the general case of Lipschitz operators for the particular instantiation of our algorithmic template.

Variance reduction Another approach to treating the stochasticity is through variance reduction techniques. There is a growing literature on this for variational inequalities [6, 18, 33, 45]. For infinite monotone operators, one approach grows the size of the mini-batches [18], which can be prohibitively expensive in large-scale settings. To this end, [33] exploits a finite sum structure and derive results for strongly monotone operators. Since this approach required knowledge of the problem constant the work, [6] instead relies on a *locally* strongly monotone structure, which is arguably more relevant for non-monotone settings faced in practice. Despite this development variation reduction techniques are known to be brittle to parameter choices possibly explaining their limited use in practice.

Relative noise The assumption of relative noise we rely on dates back to at least Polyak under the name of *relative random noise* [35]. It is a common assumption in the optimization literature but has gone under the guise of various names such as *multiplicative noise* [19]. In particular, for minimization problem it is known as the *growth condition* [5, 39, 42, 44]. This has recently gained interest [13] because of its relationship with the interpolation condition shown to hold for overparameterized models in practice.

Relative noise in online learning In the online learning literature the same noise model that we consider has been studied [25]. This particular noisy feedback model, as it is called in the community, similarly allows them to get finite time last iterate convergence also in the unconstrained setting under cocoercive with unknown constant but for a standard gradient update. However, crucially, they require the relative noise factor to vanish. We get rid of this requirement by employing an extragradient scheme with a different adaptivity, obtaining a $\mathcal{O}(1/T)$-rate for the ergodic average iterate.

B Restricted gap function

In this appendix, we discuss the basic properties of the restricted merit function $\mathrm{Gap}_{\mathcal{C}}$ introduced in (Gap). For completeness, we provide the proof of Proposition 1,which itself is an extension of a similar result by [30]:

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

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

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

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

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

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

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

This would then give

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

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

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

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

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

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

C Template inequalities

In this section we shall provide the proof of the template inequality of Proposition 2. As we already argued in the main. this energy inequality will serve as a template for deriving the method specific convergence rates in the sequel. Formally, we have. the following:

Proposition 2. *Assume that $X_t, X_{t+1/2}$ are the iterates of* (GEG) *run with a non-negative, non-increasing step-size γ_t. Then, for all $x \in \mathbb{R}^d$ the following inequality holds:*

$$\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x \right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2} \sum_{t=1}^{T} \gamma_t \left\| V_{t+1/2} - V_t \right\|_*^2 - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_t} \left\| X_{t+1/2} - X_t \right\|^2 \tag{C.1}$$

Proof. By the update rule for X_{t+1} in (GEG) we get the following:

$$\begin{aligned}
\left\langle V_{t+1/2}, X_{t+1} - x \right\rangle &= \left\langle \frac{1}{\gamma_t}\gamma_t Y_t - \frac{1}{\gamma_{t+1}}\gamma_{t+1} Y_{t+1}, X_{t+1} - x \right\rangle \\
&= \left\langle \frac{1}{\gamma_t}\gamma_t Y_t - \frac{1}{\gamma_t}\gamma_{t+1} Y_{t+1}, X_{t+1} - x \right\rangle + \left\langle \frac{1}{\gamma_t}\gamma_{t+1} Y_{t+1} - \frac{1}{\gamma_{t+1}}\gamma_{t+1} Y_{t+1}, X_{t+1} - x \right\rangle \\
&= \frac{1}{\gamma_t} \left\langle \gamma_t Y_t - \gamma_{t+1} Y_{t+1}, X_{t+1} - x \right\rangle + \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \left\langle 0 - \gamma_{t+1} Y_{t+1}, X_{t+1} - x \right\rangle. \\
&= \frac{1}{\gamma_t} \left\langle X_t - X_{t+1}, X_{t+1} - x \right\rangle + \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \left\langle 0 - X_{t+1}, X_{t+1} - x \right\rangle
\end{aligned}$$

Therefore, by developing the scalar products:

$$\langle X_t - X_{t+1}, X_{t+1} - x \rangle \quad \text{and} \quad \langle 0 - X_{t+1}, X_{t+1} - x \rangle \tag{C.2}$$

we get:

$$\begin{aligned}
\left\langle V_{t+1/2}, X_{t+1} - x \right\rangle = \frac{1}{\gamma_t} &\left[\frac{1}{2} \left\| X_{t+1} - x + X_t - X_{t+1} \right\|^2 - \frac{1}{2} \left\| X_t - X_{t+1} \right\|^2 - \frac{1}{2} \left\| X_{t+1} - x \right\|^2 \right] \\
+ \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) &\left[\frac{1}{2} \left\| X_{t+1} - x - X_{t+1} \right\|^2 - \frac{1}{2} \left\| X_{t+1} \right\|^2 - \frac{1}{2} \left\| X_{t+1} - x \right\|^2 \right]
\end{aligned} \tag{C.3}$$

which in turn yields:

$$\begin{aligned}
\left\langle V_{t+1/2}, X_{t+1} - x \right\rangle \leq \frac{1}{2\gamma_t} \left\| X_t - x \right\|^2 - \frac{1}{2\gamma_t} \left\| X_t - X_{t+1} \right\|^2 - \frac{1}{2\gamma_t} \left\| X_{t+1} - x \right\|^2 + \frac{1}{2} \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \|x\|^2 \\
- \frac{1}{2\gamma_{t+1}} \left\| X_{t+1} - x \right\|^2 + \frac{1}{2\gamma_t} \left\| X_{t+1} - x \right\|^2
\end{aligned} \tag{C.4}$$

Therefore, after rearranging,

$$\begin{aligned}\frac{1}{2\gamma_{t+1}}\left\|X_{t+1}-x\right\|^2 &\leq \frac{1}{2\gamma_t}\left\|X_t-x\right\|^2+\frac{1}{2}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|x\|^2-\left\langle V_{t+1/2},X_{t+1}-x\right\rangle-\frac{1}{2\gamma_t}\left\|X_t-X_{t+1}\right\|^2\\ &=\frac{1}{2\gamma_t}\left\|X_t-x\right\|^2+\frac{1}{2}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|x\|^2-\left\langle V_{t+1/2},X_{t+1/2}-x\right\rangle\\ &\quad+\left\langle V_{t+1/2},X_{t+1/2}-X_{t+1}\right\rangle-\frac{1}{2\gamma_t}\left\|X_t-X_{t+1}\right\|^2.\end{aligned}$$

On the other hand, by invoking the update rule of $X_{t+1/2}$ in (GEG) we have:

$$\begin{aligned}\gamma_t\left\langle V_t,X_{t+1/2}-x\right\rangle &=\left\langle X_t-X_{t+1/2},X_{t+1/2}-x\right\rangle\\ &=\frac{1}{2}\left\|X_{t+1/2}-x+X_t-X_{t+1/2}\right\|^2-\frac{1}{2}\left\|X_t-X_{t+1/2}\right\|^2-\frac{1}{2}\left\|X_{t+1/2}-x\right\|^2\\ &=\frac{1}{2}\left\|X_t-x\right\|^2-\frac{1}{2}\left\|X_t-X_{t+1/2}\right\|^2-\frac{1}{2}\left\|X_{t+1/2}-x\right\|^2,\end{aligned}\tag{C.5}$$

and after dividing with γ_t and rearranging and setting $x=X_{t+1}$

$$\frac{1}{2\gamma_t}\left\|X_t-X_{t+1/2}\right\|^2+\frac{1}{2\gamma_t}\left\|X_{t+1/2}-X_{t+1}\right\|^2+\left\langle V_t,X_{t+1/2}-X_{t+1}\right\rangle=\frac{1}{2\gamma_t}\left\|X_t-X_{t+1}\right\|^2.\tag{C.6}$$

So, combining the above, we get

$$\begin{aligned}\frac{1}{2\gamma_{t+1}}\left\|X_{t+1}-x\right\|^2 \leq \frac{1}{2\gamma_t}\left\|X_t-x\right\|^2+\frac{1}{2}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|x\|^2-\left\langle V_{t+1/2},X_{t+1/2}-x\right\rangle\\ +\left\langle V_{t+1/2},X_{t+1/2}-X_{t+1}\right\rangle-\left\langle V_t,X_{t+1/2}-X_{t+1}\right\rangle\\ -\frac{1}{2\gamma_t}\left\|X_t-X_{t+1/2}\right\|^2-\frac{1}{2\gamma_t}\left\|X_{t+1/2}-X_{t+1}\right\|^2.\end{aligned}\tag{C.7}$$

Hence, we get:

$$\begin{aligned}\frac{1}{2\gamma_{t+1}}\left\|X_{t+1}-x\right\|^2 \leq \frac{1}{2\gamma_t}\left\|X_t-x\right\|^2-\left\langle V_{t+1/2},X_{t+1/2}-x\right\rangle+\frac{1}{2}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|x\|^2\\ +\underbrace{\left\langle V_{t+1/2}-V_t,X_{t+1/2}-X_{t+1}\right\rangle-\frac{1}{2\gamma_t}\left\|X_{t+1/2}-X_{t+1}\right\|^2}_{\text{(A)}}-\frac{1}{2\gamma_t}\left\|X_t-X_{t+1/2}\right\|^2.\end{aligned}\tag{C.8}$$

Moreover, by bounding (A) from above we get:

$$\begin{aligned}&\left\langle V_{t+1/2}-V_t,X_{t+1/2}-X_{t+1}\right\rangle-\frac{1}{2\gamma_t}\left\|X_{t+1/2}-X_{t+1}\right\|^2\\ &\quad\leq\frac{1}{2}\gamma_t\left\|V_{t+1/2}-V_t\right\|_*^2+\frac{1}{2\gamma_t}\left\|X_{t+1/2}-X_{t+1}\right\|^2-\frac{1}{2\gamma_t}\left\|X_{t+1/2}-X_{t+1}\right\|^2\\ &\quad\leq\frac{1}{2}\gamma_t\left\|V_{t+1/2}-V_t\right\|_*^2.\end{aligned}\tag{C.9}$$

So, finally

$$\begin{aligned}\frac{1}{2\gamma_{t+1}}\left\|X_{t+1}-x\right\|^2 \leq \frac{1}{2\gamma_t}\left\|X_t-x\right\|^2-\left\langle V_{t+1/2},X_{t+1/2}-x\right\rangle+\frac{1}{2}\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\|x\|^2\\ +\frac{1}{2}\gamma_t\left\|V_t-V_{t+1/2}\right\|_*^2-\frac{1}{2\gamma_t}\|X_t-X_{t+1/2}\|^2\end{aligned}\tag{C.10}$$

So, after rearranging and telescoping over $t=1,\dots,T$ we get:

$$\begin{aligned}\sum_{t=1}^T\left\langle V_{t+1/2},X_{t+1/2}-x\right\rangle \leq \frac{\|X_1-x\|^2}{2\gamma_1}+\frac{\|x\|^2}{2\gamma_{T+1}}-\frac{\|x\|^2}{2\gamma_1}+\frac{1}{2}\sum_{t=1}^T\gamma_t\left\|V_t-V_{t+1/2}\right\|_*^2\\ -\frac{1}{2}\sum_{t=1}^T\frac{\|X_t-X_{t+1/2}\|^2}{\gamma_t}\end{aligned}\tag{C.11}$$

The result follows by setting $X_1=0$. □

We have the following result that will help us to deal with the "noise" martingale difference component.

Lemma C.1. *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set and $h : \mathcal{C} \to \mathbb{R}$ be a 1-strongly-convex with respect to a $\|\cdot\|$ over $\mathcal{C}$. Also, assume that $\forall x \in \mathcal{C}$, $h(x) - \min_{x\in\mathcal{C}} h(x) \leq \frac{D^2}{2}$. Then, for any martingale difference $(Z_t)_{t=1}^T \in \mathbb{R}^d$, and any random vector $x \in \mathcal{C}$, we have:*

$$\mathbb{E}\left[\left\langle \sum_{t=1}^T Z_t, x \right\rangle\right] \leq \frac{\tilde{D}}{2}\sqrt{\sum_{t=1}^T \mathbb{E}\left[\|Z_t\|_*^2\right]} \tag{C.12}$$

The proof of the above lemma could be found in [2], where they present the same result under the label Proposition B.1.

D Non-adaptive analysis

Proof of Theorem 1. Since we adopt a non-increasing step-size schedule, Proposition 2 applies to this setting. Combining this with almost sure boundedness of stochastic operators,

$$\begin{aligned}
\sum_{t=1}^T \left\langle V_{t+1/2}, X_{t+1/2} - x\right\rangle &\leq \frac{\|X\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^T \gamma_t \left\|V_{t+1/2} - V_t\right\|_*^2 - \frac{1}{2}\sum_{t=1}^T \frac{1}{\gamma_t}\left\|X_{t+1/2} - X_t\right\|^2 \\
&\leq \frac{\|x\|^2}{2}\sqrt{T+1} + \sum_{t=1}^T \gamma_t \left\|V_{t+1/2}\right\|_*^2 + \gamma_t \|V_t\|_*^2 \\
&\leq \frac{\|x\|^2}{2}\sqrt{T+1} + 2M^2\sqrt{T}.
\end{aligned}$$

By monotonicity, and the definition that $V_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$,

$$\begin{aligned}
\sum_{t=1}^T \left\langle V_{t+1/2}, X_{t+1/2} - x\right\rangle &= \sum_{t=1}^T \left\langle A(X_{t+1/2}), X_{t+1/2} - x\right\rangle + \left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle \\
&\geq \sum_{t=1}^T \left\langle A(x), X_{t+1/2} - x\right\rangle + \left\langle U_{t+1/2}, X_{t+1/2} - x\right\rangle \\
&= T\left\langle A(x), \bar{X}_T - x\right\rangle + \sum_{t=1}^T \left\langle U_{t+1/2}, X_{t+1/2} - x\right\rangle
\end{aligned}$$

Plugging this lower bound into the first expression,

$$\left\langle A(x), \bar{X}_T - x\right\rangle \leq \frac{\left(\frac{\|x\|^2}{2} + 2M^2\right)\sqrt{T+1} + \sum_{t=1}^T \left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle}{T}$$

Taking supremum over $x \in \mathcal{C}$ and finally computing expectation with respect to all randomness we obtain

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T)\right] \leq \frac{\mathbb{E}\left[\sup_{x\in\mathcal{C}}\left\{\left(\frac{\|x\|^2}{2} + 2M^2\right)\sqrt{T+1} + \underbrace{\sum_{t=1}^T \left\langle U_{t+1/2}, x\right\rangle}_{\text{(A)}} - \underbrace{\sum_{t=1}^T \left\langle U_{t+1/2}, X_{t+1/2}\right\rangle}_{\text{(B)}}\right\}\right]}{T}.$$

For term (A),

$$
\begin{aligned}
\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2}, x\right\rangle\right] &\leq \mathbb{E}\left[\max_{x\in\mathcal{C}}\left\langle\sum_{t=1}^{T}U_{t+1/2}, x\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle\sum_{t=1}^{T}U_{t+1/2}, \tilde{x}\right\rangle\right] && \text{(for some } \tilde{x}\in\mathcal{C} \text{ which attains the maximum)} \\
&= \frac{\tilde{D}}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|U_{t+1/2}\right\|_*^2\right]} && \text{((by Lemma C.1))} \\
&= \frac{\tilde{D}}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left\|U_{t+1/2}\right\|_*^2 |\mathcal{F}_{t+1/2}\right]\right]} \\
&= \frac{\tilde{D}}{2}\sigma\sqrt{T} && \text{(Bounded variance)}
\end{aligned}
$$

Also, for term (B),

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle \mid \mathcal{F}_{t+1/2}\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle\mathbb{E}\left[U_{t+1/2} \mid \mathcal{F}_{t+1/2}\right], X_{t+1/2}\right\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle 0, X_{t+1/2}\right\rangle\right] && \text{(unbiasedness of } V_{t+1/2}\text{)} \\
&= 0.
\end{aligned}
$$

Finally recognizing $\sup_{x\in\mathcal{C}}\|x\| < D$ and combining the expressions for term (A) and (B),

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T)\right] \leq \frac{\mathbb{E}\left[\sup_{x\in\mathcal{C}}\left\{\left(\frac{D^2}{2}+2M^2\right)\sqrt{T+1}+\frac{\tilde{D}}{2}\sigma\sqrt{T}\right\}\right]}{T},
$$

which concludes our derivation

$$
\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T)\right] = \mathcal{O}(1/\sqrt{T})
$$

□

Proof of Proposition 3. Directly obtained by Theorem 1 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA).

□

Proof of Theorem 2.

$$
\begin{aligned}
\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x \right\rangle &= \sum_{t=1}^{T} \left\langle V_{t+1/2} - V_t, X_{t+1/2} - X_{t+1} \right\rangle + \left\langle V_t, X_{t+1/2} - X_{t+1} \right\rangle + \left\langle V_{t+1/2}, X_{t+1} - x \right\rangle \\
&= \sum_{t=1}^{T} \|V_{t+1/2} - V_t\| \|X_{t+1/2} - X_{t+1}\| + \frac{1}{\gamma} \left\langle X_t - X_{t+1/2}, X_{t+1/2} - X_{t+1} \right\rangle + \frac{1}{\gamma} \left\langle \gamma Y_t - X_{t+1}, X_{t+1} - x \right\rangle \\
&= \sum_{t=1}^{T} \frac{\gamma}{2} \|V_{t+1/2} - V_t\|^2 + \frac{1}{2\gamma} \|X_{t+1/2} - X_{t+1}\|^2 \\
&\quad + \frac{1}{2\gamma} \left(\|X_t - X\|^2 - \|X_{t+1} - X\|^2 - \|X_t - X_{t+1/2}\|^2 - \|X_{t+1/2} - X_{t+1}\|^2 \right) \\
&= \frac{\|X_1 - x\|^2}{2\gamma} + \frac{\gamma}{2} \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma} \sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2 \\
&= \frac{\|x\|^2}{2\gamma} + \frac{\gamma}{2} \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma} \sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2,
\end{aligned}
$$

where we set $X_1 = 0$. At this point the question is how to introduce the relative noise into the analysis such that we show that the stochastic/deterministic operator norms are summable. This would enable us to achieve the anticipated 1/T rate. In other words, we want to show that

$$
\mathbb{E}\left[\sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 \right] < +\infty
$$

$$
\mathbb{E}\left[\sum_{t=1}^{T} \|A(X_t)\|^2 \right] < +\infty
$$

We take expectation with respect to all randomness and lower bound the left hand side with the norm of the operator using cocoercivity. Setting $x = x^*$, where x^* is a solution of (VI),

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle \right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle | \mathcal{F}_{t+1/2} \right] \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \mathbb{E}\left[V_{t+1/2} | \mathcal{F}_{t+1/2} \right], X_{t+1/2} - x^* \right\rangle \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle A(X_{t+1/2}) - A(x^*), X_{t+1/2} - x^* \right\rangle \right] \quad \text{(Cocoercivity)} \\
&\geq \frac{1}{L} \mathbb{E}\left[\sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 \right]
\end{aligned}
$$

Plugging this into the original expression yields

$$
\frac{1}{L} \mathbb{E}\left[\sum_{t=1}^{T} \|A(X_{t+1/2})\|^2 \right] \leq \frac{\|x^*\|^2}{2\gamma} + \frac{\gamma}{2} \sum_{t=1}^{T} \|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma} \sum_{t=1}^{T} \|X_t - X_{t+1/2}\|^2
$$

With a similar approach,

$$\mathbb{E}\left[\frac{1}{L}\sum_{t=1}^{T}\|A(X_{t+1/2})\|^2 + \frac{1}{2\gamma}\sum_{t=1}^{T}\|X_t - X_{t+1/2}\|^2\right]$$
$$\geq \mathbb{E}\left[\frac{1}{L}\sum_{t=1}^{T}\|A(X_{t+1/2})\|^2 + \frac{1}{2L^2\gamma}\sum_{t=1}^{T}\|A(X_t) - A(X_{t+1/2})\|^2\right]$$
$$\geq \mathbb{E}\left[\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\}\sum_{t=1}^{T} 2\|A(X_{t+1/2})\|^2 + 2\|A(X_t) - A(X_{t+1/2})\|^2\right]$$
$$\geq \mathbb{E}\left[\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\}\sum_{t=1}^{T}\|A(X_t)\|^2\right]$$

Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T}\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\}\|A(X_t)\|^2 + \frac{1}{L}\|A(X_{t+1/2})\|^2\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + \gamma\sum_{t=1}^{T}\|V_{t+1/2} - V_t\|^2\right]$$

We now use the relative variance in the expression on the right hand side. Relying on the towering property of expectation,

$$\mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + \gamma\sum_{t=1}^{T}\|V_{t+1/2} - V_t\|^2\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + 2\gamma\sum_{t=1}^{T}\mathbb{E}\left[\|V_{t+1/2}\|^2|\mathcal{F}_{t+1/2}\right] + \mathbb{E}\left[\|V_t\|^2|\mathcal{F}_t\right]\right]$$
$$\leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + 2\gamma c\sum_{t=1}^{T}\|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right]$$

Combining last two expressions together yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\}\left(\|A(X_t)\|^2 + \|A(X_{t+1/2})\|^2\right)\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma} + 2\gamma c\sum_{t=1}^{T}\|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right]$$

Grouping the same terms on the same side of the inequality,

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} - 2\gamma c\right)\left(\|A(X_t)\|^2 + \|A(X_{t+1/2})\|^2\right)\right] \leq \mathbb{E}\left[\frac{\|x^*\|^2}{\gamma}\right]$$

As long as $\min\left\{\frac{1}{2L}, \frac{1}{4L^2\gamma}\right\} - 2\gamma c > 0$, we show that sum of operator norms with respect to both sequences are summable.

To obtain the gap, we will decompose $V_{t+1/2}$ into the full operator plus the noise,

$$\sum_{t=1}^{T}\left\langle V_{t+1/2}, X_{t+1/2} - x\right\rangle = \sum_{t=1}^{T}\left\langle A(X_{t+1/2}), X_{t+1/2} - x\right\rangle + \sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1/2} - x\right\rangle$$
$$\geq \sum_{t=1}^{T}\left\langle A(x), X_{t+1/2} - x\right\rangle + \sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1/2} - x\right\rangle \qquad \text{(Monotonicity)}$$
$$= T\left\langle A(x), \bar{X}_{t+1/2} - x\right\rangle + \sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1/2} - x\right\rangle$$

Rearranging and incorporating into the original bound,

$$\begin{aligned}
&\langle A(x),\ \bar{X}_T - x\rangle \\
&\qquad \le \frac{1}{T}\left(\frac{\|x\|^2}{2\gamma} + \sum_{t=1}^{T}\frac{\gamma}{2}\|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma}\|X_t - X_{t+1/2}\|^2 + \left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle\right),
\end{aligned}$$

We take supremum over x to retrieve the gap function and taking expectation,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T)\right] \\
&\le \mathbb{E}\left[\sup_{x\in\mathcal{C}}\left\{\frac{1}{T}\left(\frac{\|x\|^2}{2\gamma} + \sum_{t=1}^{T}\frac{\gamma}{2}\|V_{t+1/2} - V_t\|^2 - \frac{1}{2\gamma}\|X_t - X_{t+1/2}\|^2 + \left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle\right)\right\}\right] \\
&\le \frac{1}{T}\left(\frac{D^2}{2\gamma} + \sum_{t=1}^{T}\mathbb{E}\left[\gamma\|V_{t+1/2}\|^2 + \gamma\|V_t\|^2\right] + \mathbb{E}\left[\sup_{x\in\mathcal{C}}\left\{\left\langle U_{t+1/2}, x\right\rangle\right\}\right] - \mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right]\right) \\
&\le \frac{1}{T}\left(\frac{D^2}{2\gamma} + \gamma c\underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\|A(X_{t+1/2})\|^2 + \|A(X_t)\|^2\right]}_{\text{(i)}} + \underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\sup_{x\in\mathcal{C}}\left\{\left\langle U_{t+1/2}, x\right\rangle\right\}\right]}_{\text{(ii)}} - \underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right]}_{\text{(iii)}}\right),
\end{aligned}$$

where we define that $\sup_{x\in\mathcal{C}}\|x\| \le D$ and use relative variance in the last inequality.
For term (i), we have already proven that this particular summation is finite.
For term (ii),

$$\begin{aligned}
\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2}, x\right\rangle\right] &\le \mathbb{E}\left[\max_{x\in\mathcal{C}}\left\langle\sum_{t=1}^{T}U_{t+1/2}, x\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle\sum_{t=1}^{T}U_{t+1/2}, \tilde{x}\right\rangle\right] && \text{(for some } \tilde{x}\in\mathcal{C} \text{ which attains the maximum)} \\
&= \frac{\tilde{D}}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|U_{t+1/2}\right\|_*^2\right]} && \text{((by Lemma C.1))} \\
&= \frac{\tilde{D}}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|V_{t+1/2} - A(X_{t+1/2})\right\|_*^2\right]} && \text{(unbiasedness of } V_{t+1/2}\text{)} \\
&= \frac{\tilde{D}}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|V_{t+1/2}\right\|_*^2\right]} && \text{(Towering property)} \\
&= \frac{\tilde{D}}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[c\left\|A(X_{t+1/2})\right\|_*^2\right]} < +\infty && \text{(Relative variance)}
\end{aligned}$$

Finally for term (iii),

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle \mid \mathcal{F}_{t+1/2}\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle \mathbb{E}\left[U_{t+1/2} \mid \mathcal{F}_{t+1/2}\right], X_{t+1/2}\right\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle 0, X_{t+1/2}\right\rangle\right] && \text{(unbiasedness of } V_{t+1/2}\text{)} \\
&= 0.
\end{aligned}$$

Since we have shown that either the terms are finite or 0, it immediately implies that

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\bar{X}_T)\right] = \mathcal{O}(1/T)$$

□

Proof of Proposition 4. Directly obtained by Theorem 2 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA).

□

E Adaptive analysis

In this section we shall provide the proof for (GEG) run with adaptive step-sizes for the various noise profiles.Before doing so, we shall present two key building blocks that we will use for our analysis; for both absolute and relative random noise profiles. In particular, we have:

Lemma E.1 (26, 24)**.** *For all non-negative numbers $\alpha_1, \dots \alpha_t$, the following inequality holds:*

$$\sqrt{\sum_{t=1}^{T}\alpha_t} \leq \sum_{t=1}^{T}\frac{\alpha_t}{\sqrt{\sum_{i=1}^{t}\alpha_i}} \leq 2\sqrt{\sum_{t=1}^{T}\alpha_t} \tag{E.1}$$

In order to streamline the presentation of our analysis we defer the proof Lemma E.1 to Appendix G along with several variants concerning inequalities of numerical sequences. Having this result at hand, we will start presenting our analysis with the absolute random noise setting.

Proof of Theorem 3. Recalling Proposition 2 the following inequality holds:

$$\sum_{t=1}^{T}\left\langle V_{t+1/2}, X_{t+1/2} - x\right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_t\right\|_*^2 \tag{E.2}$$

Moreover, by invoking the fact that $V_{t+1/2} = A(X_{t+1/2}) + U_{t+1/2}$ we have that:

$$\sum_{t=1}^{T}\left\langle A(X_{t+1/2}), X_{t+1/2} - x\right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_t\right\|_*^2 + \sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle \tag{E.3}$$

Now, by applying the monotonicity of A we can bound from below the (LHS) as follows:

$$\sum_{t=1}^{T}\left\langle A(x), X_{t+1/2} - x\right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_t\right\|_*^2 + \sum_{t=1}^{T}\langle U_{t+1/2}, x - X_{t+1/2}\rangle \tag{E.4}$$

So, by dividing both sides by T, taking suprema on both sides over a compact neighbourhood of a solution x^* and taking expectations:

$$\mathbb{E}\left[\sup_{x\in\mathcal{C}}\left\langle A(x), \overline{X}_T - x\right\rangle\right] \leq D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\left\|V_{t+1/2}-V_t\right\|_*^2\right] + \sum_{t=1}^{T}\mathbb{E}\left[\sup_{x\in\mathcal{C}}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right] \quad \text{(E.5)}$$

which in turn yields:

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}(\overline{X}_T)\right] \leq D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\left\|V_{t+1/2}-V_t\right\|_*^2\right] + \sum_{t=1}^{T}\mathbb{E}\left[\sup_{x\in\mathcal{C}}\langle U_{t+1/2}, x - X_{t+1/2}\rangle\right] \quad \text{(E.6)}$$

Therefore, we are left to bound from above the (RHS). We shall do this term by term: For the term $D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]$ we have:

$$D^2/2\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] = D^2/2\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2}\right] \leq D^2/2\sqrt{1+4M^2T} \quad \text{(E.7)}$$

with the second inequality being obtained by the fact that V_t is almost surely bounded for all $t = 1, 1/2, \ldots$. Moreover, for the term $\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\left\|V_{t+1/2}-V_t\right\|_*^2\right]$ we have:

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\gamma_t\left\|V_{t+1/2}-V_t\right\|_*^2\right] &= \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}(\gamma_t-\gamma_{t+1})\left\|V_{t+1/2}-V_t\right\|_*^2 + \sum_{t=1}^{T}\gamma_{t+1}\left\|V_{t+1/2}-V_t\right\|_*^2\right] \\
&\leq \frac{1}{2}\left[4M^2\mathbb{E}\left[\cdot\sum_{t=1}^{T}(\gamma_t-\gamma_{t+1})\right] + \mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t+1}\left\|V_{t+1/2}-V_t\right\|_*^2\right]\right] \\
&\leq \frac{1}{2}\left[4M^2 + \mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t+1}\left\|V_{t+1/2}-V_t\right\|_*^2\right]\right]
\end{aligned} \quad \text{(E.8)}$$

Now by applying Lemma E.1 we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t+1}\left\|V_{t+1/2}-V_t\right\|_*^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\frac{\left\|V_{t+1/2}-V_t\right\|_*^2}{\sqrt{1+\sum_{j=1}^{t}\left\|V_{j+1/2}-V_j\right\|_*^2}}\right] \\
&\leq 2\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2}\right] \\
&\leq 2\sqrt{1+4M^2T}
\end{aligned}$$

with the last inequality being obtained by the fact that V_t is bounded almost surely for all $t = 1, 1/2, \ldots$ Finally, for the term Bound (B)

$$\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle\right] = \underbrace{\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2}, x\right\rangle\right]}_{\text{(B1)}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1}\right\rangle\right]}_{\text{(B2)}} \quad \text{(E.9)}$$

For the term (B2) we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left\langle U_{t+1/2}, X_{t+1/2}\right\rangle \mid \mathcal{F}_{t+1/2}\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle \mathbb{E}\left[U_{t+1/2} \mid \mathcal{F}_{t+1/2}\right], X_{t+1/2}\right\rangle\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\left\langle 0, X_{t+1/2}\right\rangle\right] && \text{(unbiasedness of } V_{t+1/2}\text{)} \\
&= 0.
\end{aligned}$$

For the term (B1) we will use Lemma C.1 and we get:

$$\begin{aligned}
\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2}, x\right\rangle\right] &\leq \mathbb{E}\left[\max_{x\in\mathcal{C}}\left\langle \sum_{t=1}^{T} U_{t+1/2}, x\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \sum_{t=1}^{T} U_{t+1/2}, \tilde{x}\right\rangle\right] && \text{(for some } \tilde{x}\in\mathcal{C} \text{ which attains the maximum)} \\
&\leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|U_{t+1/2}\right\|_*^2\right]} && \text{(by Lemma C.1)} \\
&\leq \frac{D\sigma}{2}\sqrt{T}
\end{aligned}$$

Therefore, by combining all the above the result follows. □

Now, we can apply Theorem 3 to directly obtain Proposition 6:

Proposition 6. *Under Assumption 2 the iterates of* (DA), (DE), (OptDA) *enjoy the following:*

$$\mathbb{E}\left[Gap_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/\sqrt{T}) \tag{E.10}$$

Proof. Directly obtained by Theorem 3 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA). □

Now, we turn our attention towards the relative random noise. In particular, in order to show our main results for this context we will the following proposition as a stepping stone. As a prelude, we point out that the following result will also play a crucial role for establishing the last iterate convergence in Appendix F.

Proposition E.1. *Assume that $X_t, X_{t+1/2}$ are the iterates of* (GEG) *run with* (Adapt)*. Then, we have:*

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] = \mathbb{E}\left[1 + \sum_{t=1}^{T}\|V_t - V_{t+1/2}\|_*^2\right] < +\infty \tag{E.11}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T}\|A(X_{t+1/2})\|_*^2\right] < +\infty \tag{E.12}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T}\|A(X_t)\|_*^2\right] < +\infty \tag{E.13}$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T}\|X_{t+1/2} - X_t\|^2\right] < +\infty \tag{E.14}$$

Proof. Applying Proposition 2 and for $x = x^*$ with x^* being a solution of (VI), we have

$$\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle \leq \frac{\|x^*\|^2}{2\gamma_{T+1}} + \underbrace{\frac{1}{2}\sum_{t=1}^{T} \gamma_k \left\|V_{t+1/2} - V_t\right\|_*^2}_{\text{(A)}} - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\gamma_t} \left\|X_{t+1/2} - X_t\right\|^2 \tag{E.15}$$

First, we shall bound from above term (A):

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T} \gamma_t \left\|V_{t+1/2} - V_t\right\|_*^2 &= \frac{1}{2}\left[\sum_{t=1}^{T} (\gamma_t - \gamma_{t+1}) \left\|V_{t+1/2} - V_t\right\|_*^2 + \sum_{t=1}^{T} \gamma_{t+1} \left\|V_{t+1/2} - V_t\right\|_*^2\right] \\
&\leq \frac{1}{2}\left[4G^2 \cdot \sum_{t=1}^{T} (\gamma_t - \gamma_{t+1}) + \sum_{t=1}^{T} \gamma_{t+1} \left\|V_{t+1/2} - V_t\right\|_*^2\right] \\
&\leq 2G^2 + \frac{1}{2}\sum_{t=1}^{T} \gamma_{t+1} \left\|V_{t+1/2} - V_t\right\|_*^2 \\
&\leq 2G^2 \sqrt{1 + \sum_{t=1}^{T} \left\|V_{t+1/2} - V_t\right\|_*^2} + \frac{1}{2}\sum_{t=1}^{T} \frac{\left\|V_{t+1/2} - V_t\right\|_*^2}{\sqrt{1 + \sum_{j=1}^{t} \left\|V_{j+1/2} - V_j\right\|_*^2}} \\
&\leq 2G^2 \sqrt{1 + \sum_{t=1}^{T} \left\|V_{t+1/2} - V_t\right\|_*^2} + 2 \cdot \frac{1}{2}\sqrt{1 + \sum_{t=1}^{T} \left\|V_{t+1/2} - V_t\right\|_*^2} \\
&= \left(2G^2 + 1\right) \sqrt{1 + \sum_{t=1}^{T} \left\|V_{t+1/2} - V_t\right\|_*^2} \\
&= \left(2G^2 + 1\right) \frac{1}{\gamma_{T+1}}
\end{aligned} \tag{E.16}$$

So, the above becomes, if we also take expectations on both sides:

$$\begin{aligned}
\text{(B)} = \mathbb{E}\left[\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle\right] &\leq \frac{\|x^*\|^2}{2} \cdot \mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \left(2G^2 + 1\right) \cdot \mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\
&\quad - \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{\gamma_t} \left\|X_{t+1/2} - X_t\right\|^2\right] \\
&= \left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right] \mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] - \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{\gamma_t} \|X_{t+1/2} - X_t\|^2\right]
\end{aligned} \tag{E.17}$$

For term (B) we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\left\langle V_{t+1/2}, X_{t+1/2} - x^* \right\rangle \mid \mathcal{F}_{t+1/2}\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \mathbb{E}\left[V_{t+1/2} \mid \xi_{t+1/2}\right], X_{t+1/2} - x^* \right\rangle\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\left\langle A\left(X_{t+1/2}\right), X_{t+1/2} - x^* \right\rangle\right]
\end{aligned} \tag{E.18}$$

and since A is $1/L$-cocoercive, we get:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle V_{t+1/2}, X_{t+1/2}-x^*\right\rangle\right] \geq \sum_{t=1}^{T}\frac{1}{L}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right] \tag{E.19}$$

Therefore, by combining (E.17) and (E.19) the first inequality that we get is

$$\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right] \leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{E.20}$$

Moreover, we have:

$$\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right] \leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]-\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\gamma_t}\left\|X_{t+1/2}-X_t\right\|^2\right] \tag{E.21}$$

and after rearranging and using the fact that $1/\gamma_t \geq 1$ we have

$$(\mathrm{C}) = \frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right]+\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|X_{t+1/2}-X_t\right\|^2\right] \leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]E\left[\frac{1}{\gamma_{T+1}}\right] \tag{E.22}$$

For the term (C) we will have the following

$$\begin{aligned}
&\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right]+\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|X_{t+1/2}-X_t\ |\right\|^2\right] \\
&\qquad\geq \frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right]+\frac{1}{2L^2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)-A\left(X_t\right)\right\|_*^2\right] \\
&\qquad\geq \min\left\{\frac{1}{L},\frac{1}{2L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2+\left\|A\left(X_{t+1/2}\right)-A\left(X_t\right)\right\|_*^2\right] \\
&\qquad= \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[2\left\|A\left(X_{t+1/2}\right)\right\|_*^2+2\left\|A\left(X_{t+1/2}\right)-A\left(X_t\right)\right\|_*^2\right] \\
&\qquad\geq \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_t\right)\right\|_*^2\right]
\end{aligned} \tag{E.23}$$

So, we get the following inequalities:

$$\begin{aligned}
\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right] &\leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\
\min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_t\right)\right\|_*^2\right] &\leq \left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]
\end{aligned} \tag{ineq}$$

and

$$(\mathrm{D}) = \frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right]+\min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_t\right)\right\|_*^2\right] \leq 2\left[\frac{\|x^*\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{E.24}$$

For term (D) we have:

$$
\begin{aligned}
\frac{1}{L}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right] + \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_t\right)\right\|_*^2\right] & \\
&\geq \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\left[\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right] + \sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_t\right)\right\|_*^2\right]\right] \\
&\geq \min\left\{\frac{1}{2L},\frac{1}{4L^2}\right\}\left[\sum_{t=1}^{T}\frac{1}{c}\mathbb{E}\left[\left\|V_{t+1/2}\right\|_*^2\right] + \sum_{t=1}^{T}\frac{1}{c}\mathbb{E}\left[\left\|V_t\right\|_*^2\right]\right] \qquad \text{(Assumption 3)} \\
&\geq \frac{1}{c\max\left\{4L,8L^2\right\}}\left[\sum_{t=1}^{T}\mathbb{E}\left[2\left\|V_{t+1/2}\right\|_*^2 + 2\left\|V_t\right\|_*^2\right]\right] \\
&\geq \frac{1}{c\max\left\{4L,8L^2\right\}}\sum_{t=1}^{T}\mathbb{E}\left[\left\|V_{t+1/2}-V_t\right\|_*^2\right]
\end{aligned}
$$

So we get:

$$
\text{(E)} = \mathbb{E}\left[\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2\right] \leq 8c\max\left\{L,2L^2\right\}\cdot\left[\frac{\left\|x^*\right\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{E.25}
$$

For the term (E) we have:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2 + 1 - 1\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2 + 1\right] - 1 \\
&= \mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] - 1
\end{aligned} \tag{E.26}
$$

Therefore

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] &\leq 8c\max\left\{L,2L^2\right\}\left[\frac{\left\|x^*\right\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + 1 \\
&\leq 8c\max\left\{L,2L^2\right\}\left[\frac{\left\|x^*\right\|^2}{2}+2G^2+1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \\
&= \left[8c\max\left\{L,2L^2\right\}\left[\frac{\left\|x^*\right\|^2}{2}+2G^2+1\right]+1\right]\underbrace{\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]}_{\text{(F)}}
\end{aligned} \tag{E.27}
$$

For term (F) we have

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] &= \mathbb{E}\left[\sqrt{1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2}\right] \\
&\leq \sqrt{\mathbb{E}\left[1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2\right]} = \sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]}
\end{aligned} \tag{E.28}
$$

So, finally we get:

$$
\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \leq \left[8c\max\left\{L,2L^2\right\}\left[\frac{\left\|x^*\right\|^2}{2}+2G^2+1\right]+1\right]\sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]} \tag{E.29}
$$

Hence,

$$\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \leq \left(8c\max\left\{L, 2L^2\right\}\left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right] + 1\right)^2 \tag{E.30}$$

and the first result follows. The second and third claim is derived directly by combining the first claim with (ineq). Finally, the last summability condition by rearranging (E.21), we get:

$$\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\gamma_t}\left\|X_{t+1/2} - X_t\right\|^2\right] \leq \left[\frac{\|x^*\|^2}{2} + 2G^2 + 1\right]\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] \tag{E.31}$$

and the result by our first summability claim. □

Finally, we shall present the proof of the main result under relative random noise

Proof of Theorem 4. Recalling Proposition 2, we have for all $x \in \mathbb{R}^d$

$$\sum_{t=1}^{T}\left\langle V_{t+1/2}, X_{t+1/2} - x\right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_t\right\|_*^2 \tag{E.32}$$

By the definition of $V_{t+1/2} = A\left(X_{t+1/2}\right) + U_{t+1/2}$ the above becomes

$$\underbrace{\sum_{t=1}^{T}\left\langle A\left(X_{t+1/2}\right), X_{t+1/2} - x\right\rangle}_{\text{(A)}} + \sum_{t=1}^{T}\left\langle U_{t+1/2}, X_{t+1} - x\right\rangle \leq \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{1}\gamma_1\left\|V_{t+1} - V_t\right\|_*^2 \tag{E.33}$$

Term (A) due to monotonicity of the operator A,

$$\sum_{t=1}^{T}\left\langle A\left(X_{t+1/2}\right), X_{t+1/2} - x\right\rangle \geq \sum_{t=1}^{T}\left\langle A(x), X_{t+1/2} - x\right\rangle \text{ for all } x \in \mathbb{R}^d \tag{E.34}$$

Therefore, the above becomes after rearranging,

$$\sum_{t=1}^{T}\left\langle A(x), X_{t+1/2} - x\right\rangle \leq \sum_{t=1}^{T}\left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle + \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_t\right\|_*^2 \tag{E.35}$$

and by dividing both sides by T

$$\left\langle A(x), \bar{X}_T - x\right\rangle \leq \frac{1}{T}\left[\sum_{t=1}^{T}\left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle + \frac{\|x\|^2}{2\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_t\right\|_*^2\right] \tag{E.36}$$

and taking suprema on both sides over $\mathcal{C}$ (defining $D^2 = \sup_{x\in\mathcal{C}}\|x - x_1\|^2$),

$$\mathrm{Gap}_{\mathcal{C}}\left(\bar{X}_T\right) \leq \frac{1}{T}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle + \frac{D^2}{2\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_t\right\|_*^2\right] \tag{E.37}$$

and taking expectations:

$$\mathbb{E}\left[\mathrm{Gap}_{\mathcal{C}}\left(\bar{X}_T\right)\right] \leq \frac{1}{T}\Bigg[\underbrace{\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2}, x - X_{t+1/2}\right\rangle\right]}_{\text{(B)}} + \underbrace{\frac{D^2}{2}\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right]}_{\text{(C)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2} - V_1\right\|_*^2\right]}_{\text{(D)}}\Bigg] \tag{E.38}$$

Now, we shall bound the terms (B), (C), (D) individually. Since (B) is the most tricky one we will leave it last.

Bound (C)

$$\begin{aligned}\frac{D^2}{2}\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] &= \frac{D^2}{2}\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2}\right] \\ &\leq \frac{D^2}{2}\sqrt{\mathbb{E}\left[1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2\right.} \\ &= \frac{D^2}{2}\sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]} \\ &< +\infty, \text{ from Proposition E.1.}\end{aligned} \tag{E.39}$$

Bound (D)

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\gamma_t\left\|V_{t+1/2}-V_t\right\|_*^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\gamma_t-\gamma_{t+1}\right)\left\|V_{t+1/2}-V_t\right\|_*^2\right]+\mathbb{E}\left[\sum_{t=1}^{T}\gamma_{t+1}\left\|V_{t+1/2}-V_t\right\|_*^2\right] \\ &\leq 2G^2+2\mathbb{E}\left[\sqrt{1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2}\right] \\ &\leq 2G^2+2\sqrt{\mathbb{E}\left[1+\sum_{t=1}^{T}\left\|V_{t+1/2}-V_t\right\|_*^2\right]} \\ &\leq 2G^2+2\sqrt{\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right]} \\ &< +\infty, \text{ from Proposition E.1}\end{aligned} \tag{E.40}$$

Bound (B)

$$\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2},x-X_{t+1/2}\right\rangle\right] = \underbrace{\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2},x\right\rangle\right]}_{\text{(B1)}}-\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left\langle U_{t+1/2},X_{t+1}\right\rangle\right]}_{\text{(B2)}} \tag{E.41}$$

By working in the same spirit Theorem 3 for the term (B2) we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle U_{t+1/2},X_{t+1/2}\right\rangle\right]=0. \tag{E.42}$$

and for term (B1),

$$\mathbb{E}\left[\sup_{x\in\mathcal{C}}\sum_{t=1}^{T}\left\langle U_{t+1/2},x\right\rangle\right]\leq\frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|U_{t+1/2}\right\|_*^2\right]} \tag{E.43}$$

Due to the definition of $V_{t+1/2}=A\left(X_{t+1/2}\right)+U_{t+1/2}$ we have $U_{t+1/2}=A\left(X_{t+1/2}\right)-V_{t+1/2}$. So,

$$\begin{aligned}\mathbb{E}\left[\sup_{x\in e}\sum_{t=1}^{T}\left\langle U_{t+1/2},x\right\rangle\right] &\leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1}\right)-V_{t+1/2}\right\|_*^2\right]} \\ &\leq \frac{D}{2}\sqrt{2\sum_{t=1}^{T}\mathbb{E}\left[\left\|A\left(X_{t+1/2}\right)\right\|_*^2\right]+2\sum_{t=1}^{T}\mathbb{E}\left[\left\|V_{t+1/2}\right\|_*^2\right]} \\ &< +\infty \text{ , by Proposition E.1}\end{aligned} \tag{E.44}$$

□

Similar to Proposition 6, Theorem 4 allows us to obtain the following result.

Proposition 7. *Under Assumption 3 the iterates of* (DA)*,* (DE)*,* (OptDA) *enjoy the following:*

$$\mathbb{E}\left[Gap_{\mathcal{C}}(\overline{X}_T)\right] = \mathcal{O}(1/T) \tag{E.45}$$

Proof. Directly obtained by Theorem 4 by setting $V_t = 0$ for (DA), $V_t = g_{t+1/2}$ for (DE) and $V_t = g_{t-1/2}$ for (OptDA). □

F Last iterate analysis

We conclude by showing that the iterates $X_{t+1/2}, X_t$ of (GEG) run with the adaptive step-size policy (Adapt) converge towards some (VI) solution x^* almost surely. In doing so, we will need the following proposition:

Proposition F.1. *Let there be a non-empty closed set F and let a sequence $(x_t)_t \in \mathbb{R}^d$. Suppose that for all $z \in F$ there exists $(\beta_t)_t$ sequence of random variables satisfying the following almost surely:*

$$\mathbb{E}\left[\|x_{t+1} - z\|^2 \mid \mathcal{F}_t\right] \leq \|x_t - z\|^2 + \beta_t \tag{F.1}$$

with $\sum_{t=1}^{\infty} \beta_t < +\infty$ almost surely. Then, the following hold:

1. *$\|x_t - z\|^2$ converges almost surely.*
2. *If the set of almost sure limit points, i.e.*

$$\hat{\mathcal{X}} = \{\hat{x} \in \mathbb{R}^d : \textit{there exists a subsequence } x_{t_n} \to \hat{x} \textit{ almost surely}\} \tag{F.2}$$

is non-empty and $\hat{\mathcal{X}} \subset F$, then x_t converges almost surely to some random variable $\hat{x} \in F$.

Proof. See [8, Proposition 2.3]. □

Moreover, we will heavily use the following classical convergence theorem; that of the so-called Monotone Convergence Theorem.

Proposition F.2 (Monotone Convergence Theorem)**.** *Let (Ω, Σ, μ) be a measure space and $\mathcal{X} \in \Sigma$. Consider a pointwise non-decreasing sequence $(f_t)_t \cdot (\Sigma, \mathcal{B}_{\mathbb{R}_{>0}})$-measurable non-negative functions: $f_t : \mathcal{X} \to [0, +\infty]$. Set the pointwise limit of the (f_n),*

$$\lim_t f_t(x) = f(x) \tag{F.3}$$

Then, f is $(\Sigma, \mathcal{B}_{\mathbb{R}_{>0}})$ -measurable and

$$\lim_{t \to +\infty} \int_{\mathcal{X}} f_t d\mu = \int_{\mathcal{X}} f d\mu. \tag{F.4}$$

Having all these at hand, we are now in the position to illustrate the last iterate convergence result for the iterates of (DA)/(DE)/(OptDA). For the ease of presentation we shall provide the generic convenience of the general choice for the $V_{t+1/2}$.

Proposition F.3. *The iterates of* (DA)/(DE)/(OptDA) *converge towards a* (VI) *solution x^*.*

Proof. We are left to show that the iterates $X_{t+1/2}$ satisfies the requirements of Proposition F.1. In particular, invoking Proposition 2 we have:

$$\frac{1}{2\gamma_{t+1}} \|X_{t+1} - x^*\|^2 \leq \frac{1}{2\gamma_t} \|X_t - x^*\|^2 - \langle V_{t+1/2}, X_{t+1/2} - x^* \rangle + \frac{D^2}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right) + \gamma_t \left\|V_t - V_{t+1/2}\right\|_*^2 \tag{F.5}$$

with $D^2 = \sup_{x^* \in \mathcal{X}} \|x^*\|^2$. Now, by multiplying both sides with $2\gamma_t$ and using the fact that γ_t is non-decreasing and $\gamma_t \leq 1$ we get:

$$\|X_{t+1} - x^*\|^2 \leq \|X_t - x^*\| - \gamma_t \langle V_{t+1/2}, X_{t+1/2} - x^* \rangle + \frac{D^2}{2}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right) + \left\|V_t - V_{t+1/2}\right\|_*^2 \tag{F.6}$$

Now, by taking conditional expectations we obtain:

$$\mathbb{E}\left[\|X_{t+1}-x^*\|^2 \mid \mathcal{F}_{t+1/2}\right] \leq \|X_t - x^*\|^2 - \gamma_t \mathbb{E}\left[\langle V_{t+1/2}, X_{t+1/2}-x^*\rangle \mid \mathcal{F}_{t+1/2}\right] + \frac{D^2}{2}\mathbb{E}\left[\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right) \mid \mathcal{F}_{t+1/2}\right] + \gamma_t \mathbb{E}\left[\|V_t - V_{t+1}\|_*^2 \mid \mathcal{F}_{t+1/2}\right] \quad \text{(F.7)}$$

since γ_t is $\mathcal{F}_{t+1/2}$- measurable. Moreover, we have:

$$\gamma_t \mathbb{E}\left[\langle V_{t+1/2}, X_{t+1/2}-x^*\rangle \mid \mathcal{F}_{t+1/2}\right] = \gamma_t \langle \mathbb{E}[V_{t+1/2}|\mathcal{F}_{t+1/2}], X_{t+1/2}-x^*\rangle \leq 0 \quad \text{(F.8)}$$

since x^* is a solution of (VI) and $V_{t+1/2}$ is an unbiased estimator of $A(X_{t+1/2})$. So, we obtain:

$$\mathbb{E}\left[\|X_{t+1}-x^*\|^2 \mid \mathcal{F}_{t+1/2}\right] \leq \|X_t - x^*\|^2 + \frac{D^2}{2}\mathbb{E}\left[\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right) \mid \mathcal{F}_{t+1/2}\right] + \mathbb{E}\left[\|V_t - V_{t+1}\|_*^2 \mid \mathcal{F}_{t+1/2}\right] \quad \text{(F.9)}$$

The first step is to show that:

$$\beta_t = \frac{D^2}{2}\mathbb{E}\left[\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right) \mid \mathcal{F}_{t+1/2}\right] + \mathbb{E}\left[\|V_t - V_{t+1}\|_*^2 \mid \mathcal{F}_{t+1/2}\right] \quad \text{(F.10)}$$

Indeed we have that:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \beta_t\right] &= \frac{D^2}{2}\mathbb{E}\left[\sum_{t=1}^T\left(\frac{1}{\gamma_{t+1}}-\frac{1}{\gamma_t}\right)\right] + \mathbb{E}\left[\sum_{t=1}^T \|V_t - V_{t+1}\|_*^2\right] \\
&\leq \frac{D^2}{2}\mathbb{E}\left[\frac{1}{\gamma_{T+1}}\right] + \mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \\
&\leq \left(\frac{D^2}{2}+1\right)\mathbb{E}\left[\frac{1}{\gamma_{T+1}^2}\right] \\
&< +\infty
\end{aligned}$$

due to Proposition E.1. On the other hand, $\sum_{t=1}^T \beta_t$ is a non-decreasing (random) sequence; therefore is converges almost surely to some random value $\sum_{t=1}^{+\infty} \beta_t \in (0, \infty]$. Assume that $\beta_\infty = +\infty$. Then, by applying Proposition F.2 we get:

$$+\infty = \mathbb{E}\left[\sum_{t=1}^{+\infty}\beta_t\right] = \lim_T \mathbb{E}\left[\sum_{t=1}^T \beta_t\right] < +\infty \quad \text{(F.11)}$$

which is a contradiction. Therefore $\sum_{t=1}^{+\infty} \beta_t < +\infty$ almost surely. Therefore, we are left to show that every almost sure limit point of X_t is a (VI) solution. Let $\hat{x} \in \mathbb{R}^d$ be a limit point of X_t. Then, there exists a subsequence X_{t_n} which converges almost surely towards $\hat{x}$. Then, by invoking Proposition E.1 (ii), we have that:

$$\mathbb{E}\left[\sum_{t=1}^T \|A(X_t)\|_*^2\right] < +\infty \quad \text{(F.12)}$$

Therefore by the same reasoning as above, Proposition F.2 ensures that:

$$\sum_{t=1}^T \|A(X_t)\|_*^2 < +\infty \ \text{ almost surely} \quad \text{(F.13)}$$

which yields a fortiori that $\|A(X_t)\|_*^2 \to 0$ almost surely. On the other hand, we have that: $\|A(X_{t_n})\|_* \to \|A(\hat{x})\|_*$. Thus, by limit uniqueness we get that that $\|A(\hat{x})\|_* = 0$, so $\hat{x}$ is a (VI) solution, hence the result follows by Proposition F.1. Finally, in order to show that $X_{t+1/2}$ converges also towards a solution, we shall invoke Proposition E.1 (iii) that:

$$\mathbb{E}\left[\sum_{t=1}^T \|X_t - X_{t+1/2}\|^2\right] < +\infty \quad \text{(F.14)}$$

Hence, by the same reasoning we obtain that:

$$\|X_t - X_{t+1/2}\|^2 \to 0 \ \text{ almost surely} \quad \text{(F.15)}$$

and so our proof is completed. □

G Lemmas on numerical sequences

In this appendix, we provide the necessary inequality on numerical sequences that we require for the convergence rate analysis of the previous sections.

This lemma that we present is due to [26] and [24].

Lemma E.1 (26, 24)**.** *For all non-negative numbers $\alpha_1, \dots \alpha_t$, the following inequality holds:*

$$\sqrt{\sum_{t=1}^{T} \alpha_t} \leq \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} \leq 2\sqrt{\sum_{t=1}^{T} \alpha_t} \tag{G.1}$$

Proof. We begin, by introducing some necessary notation and set $S = \sum_{t=1}^{T} \alpha_t$ and $x = \alpha_T$.

The first part is proved by induction. The induction base case $T = 1$ straightforwardly holds. Now for the induction step, assume that the lemma holds for $T - 1$ and we will show that it holds for T as well. In particular, we have:

$$\begin{aligned} \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} &= \sum_{t=1}^{T-1} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} + \frac{\alpha_T}{\sqrt{\sum_{t=1}^{T} \alpha_t}} \\ &\geq \sqrt{\sum_{t=1}^{T-1} \alpha_t} + \frac{\alpha_T}{\sqrt{\sum_{t=1}^{T} \alpha_t}} = \sqrt{S - x} + \frac{x}{\sqrt{S}} \end{aligned} \tag{G.2}$$

where the first inequality is obtained due to the induction hypothesis. Hence, in order to prove the lemma it is sufficient to show that:

$$\sqrt{S - x} + \frac{x}{\sqrt{S}} \geq \sqrt{S} \tag{G.3}$$

By multiplying both sides by $\sqrt{S}$, we get the following equivalent expression:

$$\sqrt{S^2 - xS} \geq S - x \tag{G.4}$$

whereas after rearranging we obtain the equivalent inequality:

$$x \leq S \tag{G.5}$$

which holds, since $x = \alpha_T \leq \sum_{t=1}^{T} \alpha_t = S$ and hence the LHS inequality is obtained.
Now, the proof of the RHS inequality:

$$\sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} \leq 2\sqrt{\sum_{t=1}^{T} \alpha_t} \tag{G.6}$$

will again be done again via induction. The induction base $T = 1$ holds immediately. Assume that the lemma holds for $T - 1$. We will show that it also holds for T. By the induction hypothesis, we get:

$$\sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} \leq 2\sqrt{\sum_{t=1}^{T-1} \alpha_t} + \frac{\alpha_T}{\sqrt{\sum_{t=1}^{T} \alpha_t}} = 2\sqrt{S - x} + \frac{x}{\sqrt{S}} \tag{G.7}$$

where $x = \alpha_T$ and $S = \sum_{t=1}^{T} \alpha_t$ (we highlight once more that $x \leq S$). Taking the derivative of the function $H(x) = 2\sqrt{S - x} + \frac{x}{\sqrt{S}}$ (with respect to x), we get that:

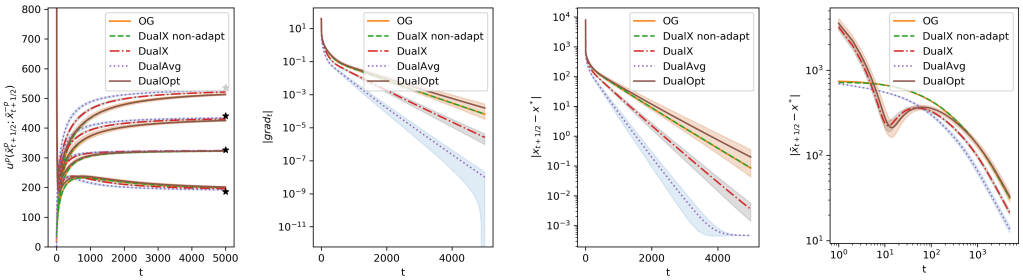
$$H'(x) = \frac{1}{\sqrt{S}} - \frac{1}{\sqrt{S - x}} \tag{G.8}$$

becomes negative for all $x \geq 0$. Thus, $H(x) \leq H(0) = 2\sqrt{S}$ and therefore the result follows. □

H Numerics

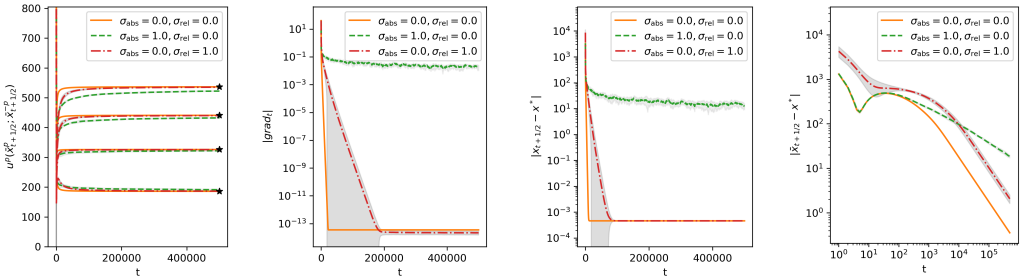
All experiments were performed on a MacBook Pro with a 2.7 GHz Quad-Core Intel Core i7 processor and 16 GB of RAM.

H.1. Kelly auction. We consider a Kelly auction with number of player $N = 4$, cost of bidding $Z = 100$, resources $Q = 1000$ and marginal utility gains $G = (1.8, 2.0, 2.2, 2.4)$ (see Section 6 for exact definitions). The hyperparameters (the step-size for non-adaptive methods and the diameter for adaptive methods) are not fine-tuned but chosen heuristically based on the sweep in Fig. H.5. When error bars are present they represent one standard deviation based on 10 independent executions. For more information on naming and notation see Section 6.



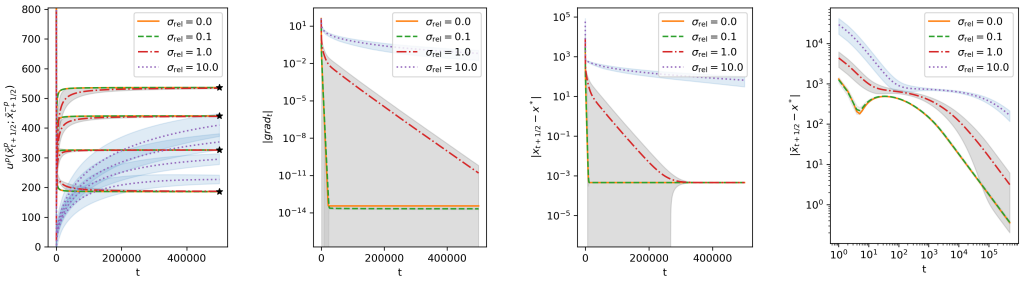
(a) Player utility **(b)** Gradient norm **(c)** Last iterate **(d)** Average iterate

Figure H.1: Comparing methods with $\sigma_{\text{rel}} = 0.1$.



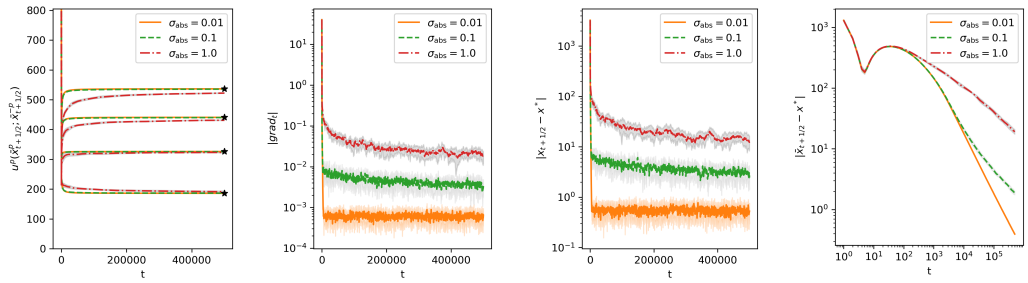
(a) Player utility **(b)** Gradient norm **(c)** Last iterate **(d)** Average iterate

Figure H.2: Relative noise compared with absolute noise using adaptive DualX. We observe the deterioration of the rate for the average iterate for absolute noise in contrast with relative noise.



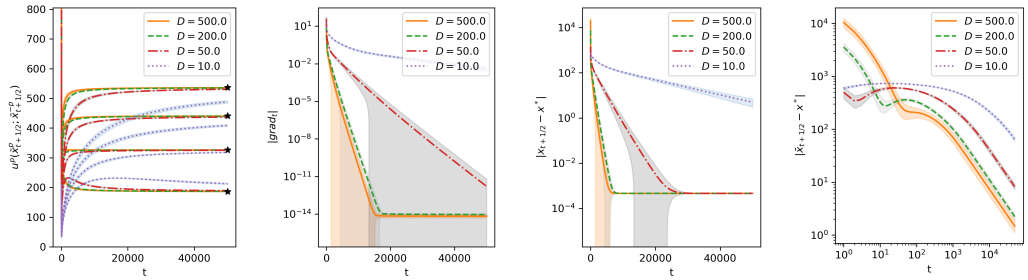
(a) Player utility **(b)** Gradient norm **(c)** Last iterate **(d)** Average iterate

Figure H.3: Different levels of relative noise using adaptive DualX. The $\mathcal{O}(1/T)$ rate for the average iterate (last plot) is kept even when the noise level is increased to $\sigma_{\text{rel}} = 1.0$.



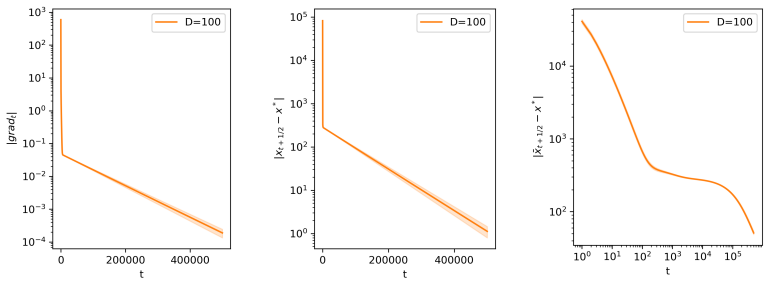
(a) Player utility **(b)** Gradient norm **(c)** Last iterate **(d)** Average iterate

Figure H.4: Different levels of absolute noise using adaptive DualX. In contrast with relative noise increasing the absolute noise clearly worsens the slope for the average iterate (last plot) indicating a rate of $\mathcal{O}(1/\sqrt{T})$.



(a) Player utility **(b)** Gradient norm **(c)** Last iterate **(d)** Average iterate

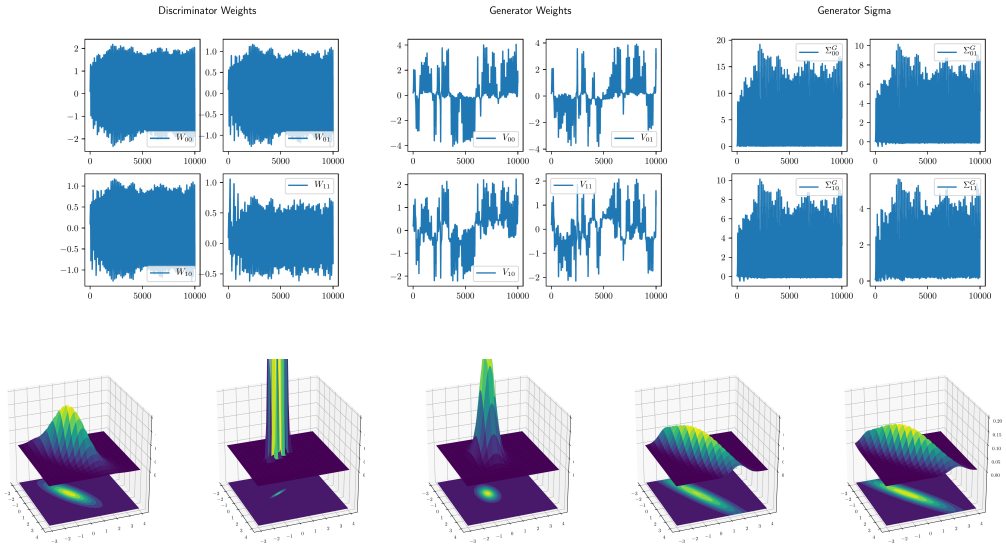
Figure H.5: We compare the effect of the diameter choice for adaptive DualX on a Kelly auction with $\sigma_{\text{rel}} = 0.2$. We can observe the $\mathcal{O}(1/T)$ average iterate rate for the whole spectrum of diameters but we note that the choice of diameter still has practical impact on convergence time. The fastest convergence is achieved with the highest diameter but note that the method did not converge for $D = 1000$ which we have excluded to keep the plots readable.



(a) Gradient norm **(b)** Last iterate **(c)** Average iterate

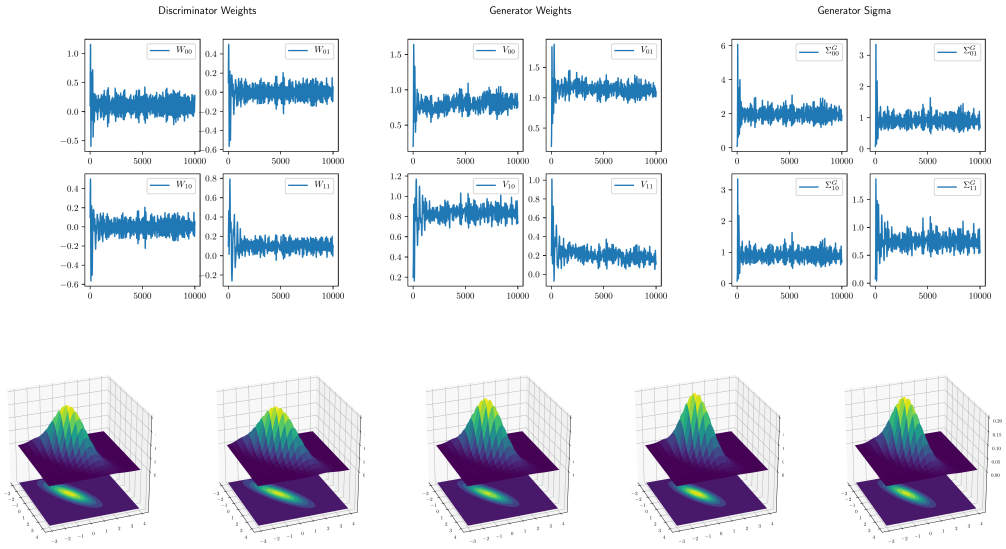
Figure H.6: DualX in a higher dimensional kelly auction ($N = 100$). Let the total resources be $Q = 1000$, cost of biddding be $Z = 100$ and the marginal utility gains be $G = (6.001, 6.002, ..., 6.1)$ (see Section 6 for exact definitions). Since the problem size is out of scope for Mathematica to provide a numerical solution we computed the optimal point using adaptive DualX for 2.5 million full steps with fine-tuned diameter in a deterministic setting. The experiment is then subsequently performed with $\sigma_{\text{rel}} = 0.1$ for 500 000 iterations.

H.2. Learning a covariance matrix. We apply adaptive DualX to the non-convex problem of learning a covariance matrix introduced in [9] (see Fig. H.9). To fit our unconstrained setting we avoid weight clipping. Thus for fair comparison we include trajectories of GD and OG as well, under these different conditions (see Fig. H.7 and Fig. H.8 respectively). The experiments builds on the code provided by the authors in under the MIT license [9].



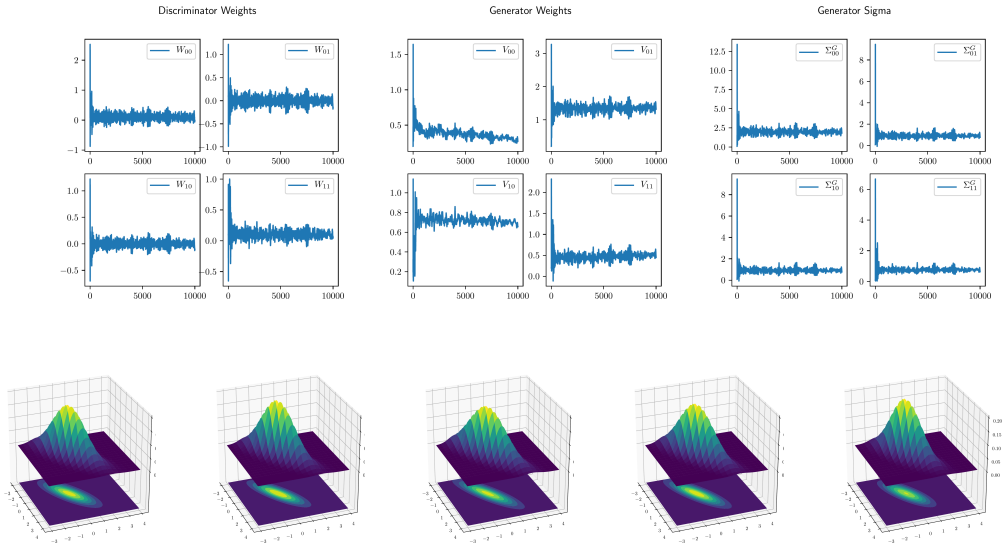
(a) True Distribution **(b)** Iterate $T - 70$ **(c)** Iterate $T - 35$ **(d)** Iterate $T - 20$ **(e)** Iterate T

Figure H.7: GD for covariance learning of a two-dimensional gaussian without weight clipping using a batch size of 50. Comparison of true distribution and distribution of generator at various points at the end of training.



(a) True Distribution **(b)** Iterate $T - 70$ **(c)** Iterate $T - 35$ **(d)** Iterate $T - 20$ **(e)** Iterate T

Figure H.8: OG for covariance learning of a two-dimensional gaussian without weight clipping using a batch size of 50. Comparison of true distribution and distribution of generator at various points at the end of training.



(a) True Distribution **(b)** Iterate $T - 70$ **(c)** Iterate $T - 35$ **(d)** Iterate $T - 20$ **(e)** Iterate T

Figure H.9: Adaptive DualX for covariance learning of a two-dimensional gaussian without weight clipping using a batch size of 50. Comparison of true distribution and distribution of generator at various points at the end of training.