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

Anonymous Author(s)
Affiliation
Address
email

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, 11].

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 [14], multi-agent and robust reinforcement learning [32], auction theory [39], 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 [21, 25] with Polyak-Rupert averaging [34]. 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 $\mathcal{O}(1/\sqrt{T})$ [19], and this rate is, in general, not improvable [27] without additional assumptions.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

	V_t	Lipschitz		Cocoercive + relative noise	
		Ergodic	Last Iterate	Ergodic	Last Iterate
Adaptive dual averaging	0	$1/\sqrt{T}$ [9]	Unknown	$1/T$	Asymptotic
Adaptive dual extrapolation	$AX_t + \text{rel.noise}$	$1/\sqrt{T}$ [36]	Unknown	$1/T$	Asymptotic
Adaptive optimistic gradient	$AX_{t-1/2} + \text{rel.noise}$	$1/\sqrt{T}$ [10]	Unknown	$1/T$	Asymptotic

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.

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 [15]; 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 [33], 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 [30], representation learning [44], and multi-agent learning [23]. This model has also been studied under the umbrella of multiplicative noise [18] or growth conditions [5, 37, 40, 42], and it is known to improve the convergence rate of stochastic gradient algorithms with non-adaptive step-sizes, even in non-smooth problems [12].

With regard to the second question, we introduce a flexible first-order algorithmic template that includes as special cases the dual averaging [29], dual extrapolation [28] and optimistic gradient methods [35, 36], 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 [23].

Our work closes several open threads in [23], 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.

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 [11, 38] 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

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

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 [28]. 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 [27]. 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:

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

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, 20, 22] and we shall denote them as *absolutely random*. On the other hand, we shall also consider the so-called *relatively random* [33]. 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 [19, 26] 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 [29, 41]. This idea is fairly standard in the literature of extra-gradient methods [21, 25, 28]. 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 [29]: 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 [28]: 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 [35, 36]: 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:*

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

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:

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

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

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

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

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

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 $\mathcal{O}(1/\sqrt{T})$ [**?**], 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:*

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

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:

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

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 [36]; 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 [36]. 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, [9, 24]:

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

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, 36].

3. (Adaptive Stochastic Optimistic Dual Averaging):. For $V_t = g_{t-1/2}$ (Adapt) becomes the step-size used in [16]:

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

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:

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

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:*

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 [13] 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

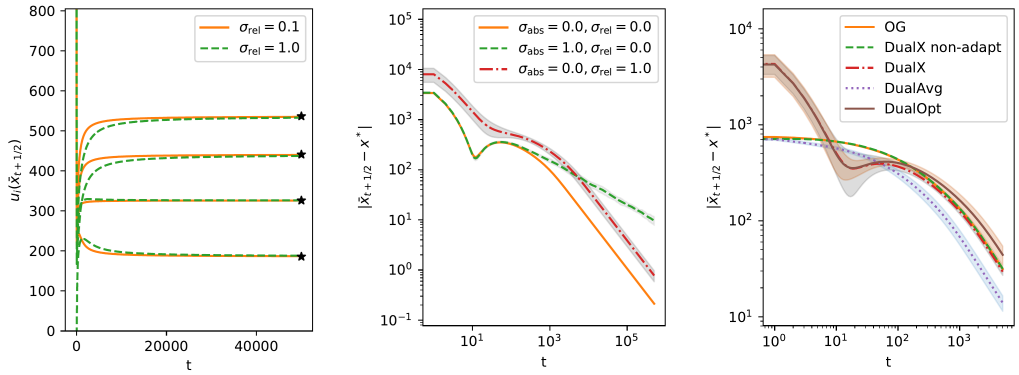

Figure 1: (left) Player utility using adaptive DualX for various relative noise levels. Even at relative high levels of noise do we converge to the optimal depicted with ($\star$). (center) Average iterate for deterministic, absolute noise and relative noise using adaptive DualX. We observe the $\mathcal{O}(1/\sqrt{T})$ 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. E.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.

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 [36] 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 [8] for comparison. For higher dimensional experiments see Appendix E.1. In Appendix E.2 we additionally apply our adaptive method to the non-convex problem of learning a covariance matrix introduced in [8].

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.

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] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

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

[30] 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.

[31] 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.

[32] 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.

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

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

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

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

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 1.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is a theoretical paper.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We plan to provide the code with the final version.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6 and Appendix E.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix E.2
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We consider simulations not data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]