A Zeroth-Order Momentum Method for Risk-Averse Online Convex Games

Zifan Wang, Yi Shen, Zachary I. Bell, Scott Nivison, Michael M. Zavlanos, and Karl H. Johansson

***Abstract*— We consider risk-averse learning in repeated unknown games where the goal of the agents is to minimize their individual risk of incurring significantly high cost. Specifically, the agents use the conditional value at risk (CVaR) as a risk measure and rely on bandit feedback in the form of the cost values of the selected actions at every episode to estimate their CVaR values and update their actions. A major challenge in using bandit feedback to estimate CVaR is that the agents can only access their own cost values, which, however, depend on the actions of all agents. To address this challenge, we propose a new risk-averse learning algorithm with momentum that utilizes the full historical information on the cost values. We show that this algorithm achieves sub-linear regret and matches the best known algorithms in the literature. We provide numerical experiments for a Cournot game that show that our method outperforms existing methods.**

I. Introduction

In online convex games [1], [2], agents interact with each other in the same environment in order to minimize their individual cost functions. Many applications, including traffic routing [3] and online marketing [4], among many others [5], can be modeled as online convex games. The individual cost functions depend on the joint decisions of all agents and are typically unknown but can be accessed via querying. Using this information, the agents can sequentially select the best actions that minimize their expected accumulated costs. This can be done using online learning algorithms whose performance is typically measured using the notions of regret [5] that quantify the gap between the agents' online decisions and the best decisions in hindsight. Many such online learning algorithms have been recently proposed, and have been shown to achieve sub-linear regret, which indicates that optimal decisions are eventually selected; see e.g., [1], [2], [6]. In this paper, here we focus on online convex games for high-stakes applications, where decisions that minimize the expected cost functions are not necessarily desirable. For example, in portfolio management, a selection of assets with the highest expected return is not necessarily desirable since it may have high volatility that can lead to catastrophic losses. To capture unexpected and catastrophic outcomes, risk-averse criteria have been widely proposed in place of the expectation, such as the Sharpe Ratio [7] and Conditional Value at Risk (CVaR) [8].

In this paper, we consider online convex games with risk-averse agents that employ CVaR as risk measure. We assume that only cost function evaluations at selected decisions are accessible. Despite the practical utility of CVaR as a measure of risk in high-stakes applications, the theoretical analysis of risk-averse learning methods that employ the CVaR is limited, mainly since the CVaR values must be estimated from the distributions of the unknown cost functions. To avoid this problem, [9] proposes a variational definition of CVaR that allows to reformulate the computation of CVaR values into an optimization problem by introducing additional variables. Building on this reformulation, many risk-averse learning methods have been proposed, including [10], [11], [12], and their convergence rates have been analysed. Specifically, [10] provides performance guarantees for stochastic gradient descent-type algorithms in risk-averse statistical learning problems; [11] addresses risk-averse online convex optimization problems with bandit feedback; and [12] proposes an adaptive sampling strategy for CVaR learning, and transforms the supervised learning problem into a zero-sum game between the sampler and the learner. However, all the above works focus on single-agent learning problems.

To the best of our knowledge, multi-agent risk-averse learning problems have not been analyzed in the literature, with only a few exceptions [13], [14]. The work in [13] solves multi-armed bandit problems with finite actions, which are different from the convex games with continuous actions considered here. Most closely related to this paper is our recent work in [14] proposing a new risk-averse learning algorithm for online convex games, where the agents utilize the cost evaluations at each time step to update their actions and achieve no-regret learning with high probability. To improve the learning rate of this algorithm, [14] also proposes two variance reduction methods on the CVaR estimates and gradient estimates, respectively. In this paper, we build on this work to significantly improve the performance of the algorithm using a notion of momentum that combines these variance reduction techniques. Specifically, the proposed new algorithm reuses samples of the cost functions from the full history of data while appropriately decreasing the weights of out-of-date samples. The idea is that, using more samples, the unknown distribution of the cost functions needed to estimate the CVaR values can be better estimated. To do so, at each time step, the proposed algorithm estimates the distribution of the cost functions using both the immediate empirical

*This work is supported in part by AFOSR under award #FA9550-19-1-0169 and by NSF under award CNS-1932011.

Zifan Wang and Karl H. Johansson are with Division of Decision and Control Systems, School of Electrical Enginnering and Computer Science, KTH Royal Institute of Technology, and also with Digital Futures, SE-10044 Stockholm, Sweden. Email: {zifanw,kallej}@kth.se.

Yi Shen and Michael M. Zavlanos are with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA. Email: {yi.shen478, michael.zavlanos}@duke.edu

Zachary I. Bell and Scott Nivison are with the Air Force Research Laboratory, Eglin AFB, FL, USA. Email: {scott.nivison, zachary.bell.10}@us.af.mil.

distribution estimate and the distribution estimate from the previous time step, which summarizes all past cost values. Then, to construct the gradient estimates, more weight is allocated to more recent information, similar to momentum gradient descent methods [15]. To reduce the variance of the gradient estimate, we employ residual feedback [16], which uses previous gradient estimates to update the current one.

Although [14] showed that reusing samples from the last iteration or residual feedback can individually improve the performance of the algorithm, it also identified that the combination of sample reuse and residual feedback is an unexplored and theoretically nontrivial problem. In this work, we combine sample reuse and residual feedback by proposing a novel zeroth-order momentum method that reuses all historical samples, compared to samples from only the last iteration. An additional important contribution of this work is that the proposed zeroth-order momentum method subsumes Algorithm 3 in [14] as a special case by appropriately choosing the momentum parameter. We show that our proposed method theoretically matches the best known result in [14] but empirically outperforms other existing methods.

The rest of the paper is organized as follows. In Section II, we formally define the problem and provide some preliminary results. In Section III, we summarize results in [14] that are needed to develop our proposed method. In Section IV, we present our proposed method and analyze its regret. In Section V, we numerically verify our method using a Cournot game example. Finally, in Section VI, we conclude the paper.

II. Problem definition

We consider a repeated game $\mathcal{G}$ with N risk-averse agents. Each agent selects an action x_i from the convex set $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$, and then receives a stochastic cost $J_i(x_i, x_{-i}, \xi_i) : \mathcal{X} \times \Xi_i \to \mathbb{R}$, where x_{-i} represents all agents' actions except for agent i, and $\mathcal{X} = \Pi_{i=1}^N \mathcal{X}_i$ is the joint action space. We sometimes instead write $J_i(x, \xi_i)$ for ease of notation, where $x = (x_i, x_{-i})$ is the concatenated vector of all agents' actions. The cost $J_i(x_i, x_{-i}, \xi_i)$ is stochastic, as $\xi_i \in \Xi_i$ is a stochastic variable. We assume that the diameter of the convex set $\mathcal{X}_i$ is bounded by D_x for all $i = 1, \ldots, N$. In what follows, we make the following assumptions on the class of games we consider in this paper.

Assumption 1. *The function $J_i(x_i, x_{-i}, \xi_i)$ is convex in x_i for every $\xi_i \in \Xi_i$ and bounded by U, i.e., $|J_i(x, \xi_i)| \leq U$, for all $i = 1, \ldots, N$.*

Assumption 2. *$J_i(x, \xi_i)$ is L_0-Lipschitz continuous in x for every $\xi_i \in \Xi_i$, for all $i = 1, \ldots, N$.*

Assumptions 1 and 2 are common assumptions in the analysis of online learning in games [17].

The goal of the risk-averse agents is to minimize the risk of incurring significantly high costs, for possibly different risk levels. In this work, we utilize CVaR as the risk measure. For a given risk level $\alpha_i \in [0, 1]$, CVaR is defined as the average of the α_i percent worst-case cost. Specifically, we denote by $F_i^x(y) = \mathbb{P}\{J_i(x, \xi_i) \leq y\}$ as the cumulative distribution function (CDF) of the random cost $J_i(x, \xi_i)$ for agent i; and by J^{α_i} the cost value at the $1-\alpha_i$ quantile of the distribution, also called the Value at Risk (VaR). Then, for a given risk level $\alpha_i \in [0, 1]$, CVaR of the cost function $J_i(x, \xi_i)$ of agent i is defined as

$$\begin{aligned} C_i(x) &:= \mathrm{CVaR}_{\alpha_i}[J_i(x, \xi_i)] \\ &:= \mathbb{E}_F[J_i(x, \xi_i) | J_i(x, \xi_i) \geq J^{\alpha_i}]. \end{aligned}$$

Notice that the CVaR value is determined by the distribution function $F_i^x(y)$ for a given α_i, so we sometimes write CVaR as a function of the distribution function, i.e., $\mathrm{CVaR}_{\alpha_i}[F_i^x] := \mathrm{CVaR}_{\alpha_i}[J_i(x, \xi_i)]$. In addition, we assume that the agents cannot observe other agents' actions, but the only information they can observe is the cost evaluations of the jointly selected actions at each time step.

Given Assumptions 1 and 2, the following lemma lists properties of CVaR, that will be important in the analysis that follows. The proof can be found in [11].

Lemma 1. *Given Assumptions 1 and 2, we have that $C_i(x_i, x_{-i})$ is convex in x_i and L_0-Lipschitz continuous in x, for all $i = 1, \ldots, N$.*

The following additional assumption on the variation of the CDFs is needed for the analysis that follows.

Assumption 3. *Let $F_i^w(y) = \mathbb{P}\{J_i(w, \xi_i) \leq y\}$ and $F_i^v(y) = \mathbb{P}\{J_i(v, \xi_i) \leq y\}$. There exist a constant $C_0 > 0$ such that*

$$\sup_y |F_i^w(y) - F_i^v(y)| \leq C_0 \|w - v\|.$$

Assumption 3 states that the difference between two CDFs can be bounded by the difference between the corresponding two action profiles. It is less restrictive than Assumption 3 in [14] since it does not depend on the algorithm but only on the cost functions.

A common measure of the ability of the agents to learn online optimal decisions that minimize their individual risk-averse objective functions $C_i(x)$ is the algorithm regret, which is defined as the difference between the actual rewards and best rewards that the agent could have achieved by playing the single best action in hindsight. Suppose the action sequences of agent i and the other agents in the team are $\{\hat{x}_{i,t}\}_{t=1}^T$ and $\{\hat{x}_{-i,t}\}_{t=1}^T$, respectively. Then, we define the regret of agent i as

$$\mathrm{R}_{C_i}(T) = \sum_{t=1}^T C_i(\hat{x}_{i,t}, \hat{x}_{-i,t}) - \min_{\tilde{x}_i \in \mathcal{X}_i} \sum_{t=1}^T C_i(\tilde{x}_i, \hat{x}_{-i,t}).$$

An algorithm is said to be no-regret if its regret grows sub-linearly with the number of episodes T. In this paper, we propose a no-regret learning algorithm, so that $\lim_{T\to\infty} \frac{\mathrm{R}_{C_i}(T)}{T} = 0$, $i = 1, \ldots, N$.

III. Preliminary Results

In this section, we summarize some results from [14] that lay the foundation for the subsequent analysis. Note that the

CVaR values depend on the distributions of the cost functions, which are generally unknown. To estimate the distribution of the cost functions, and subsequently the CVaR values with only a few samples, a sampling strategy is proposed in [14] that uses a decreasing number of samples with the number of iterations. Specifically, at episode t, the sampling strategy used by the agents is designed as follows

$$n_t = \lceil bU^2(T-t+1)^a \rceil, \tag{1}$$

where $\lceil \cdot \rceil$ is the regularized ceiling function, T is the number of episodes, U is the bound of J_i as in Assumption 1, and $a, b \in (0,1)$ are parameters to be selected later.

Using the cost evaluations at each episode, zeroth-order methods can be employed to estimate the CVaR gradients and then update the agents' actions. Specifically, at episode t, the agents calculate the number of samples n_t according to (1). Then, each agent perturbs the current action $x_{i,t}$ by an amount $\delta u_{i,t}$, where $u_{i,t} \in \mathbb{S}^{d_i}$ is a random perturbation direction sampled from the unit sphere $\mathbb{S}^{d_i} \subset \mathbb{R}^{d_i}$ and δ is the size of this perturbation. Next the agents play their perturbed actions $\hat{x}_{i,t} = x_{i,t} + \delta u_{i,t}$ for n_t times, and obtain n_t cost evaluations which are utilized to update their actions. To facilitate the theoretical analysis, we define the δ-smoothed function $C_i^\delta(x) = \mathbb{E}_{w_i \sim \mathbb{B}_i, u_{-i} \sim \mathbb{S}_{-i}}[C_i(x_i + \delta w_i, x_{-i} + \delta u_{-i})]$, where $\mathbb{S}_{-i} = \Pi_{j \neq i} \mathbb{S}_j$, and $\mathbb{B}_i$, $\mathbb{S}_i$ denote the unit ball and unit sphere in $\mathbb{R}^{d_i}$, respectively. The size of the perturbation δ here is related to a smoothing parameter that controls how well $C_i^\delta(x)$ approximates $C_i(x)$. As shown in [14], the function $C_i^\delta(x)$ satisfies the following properties.

Lemma 2. *Let Assumptions 1 and 2 hold. Then we have that*

1) $C_i^\delta(x_i, x_{-i})$ *is convex in* x_i,
2) $C_i^\delta(x)$ *is* L_0*-Lipschitz continuous in* x,
3) $|C_i^\delta(x)$-$C_i(x)| \leq \delta L_0 \sqrt{N}$,
4) $\mathbb{E}[\frac{d_i}{\delta} C_i(\hat{x}_t) u_{i,t}] = \nabla_i C_i^\delta(x_t)$.

The last property in Lemma 2 shows that the term $\frac{d_i}{\delta} C_i(\hat{x}_t) u_{i,t}$ is an unbiased estimate of the gradient of the smoothed function $C_i^\delta(x)$. Next we present a lemma that helps bound the distance between two CVaR values by the distance between the two corresponding CDFs. The proof can be found in [14].

Lemma 3. *Let* F *and* G *be two CDFs of two random variables and suppose the random variables are bounded by* U*. Then we have that*

$$|\mathrm{CVaR}_\alpha[F] - \mathrm{CVaR}_\alpha[G]| \leq \frac{U}{\alpha} \sup_y |F(y) - G(y)|.$$

IV. A ZEROTH-ORDER MOMENTUM METHOD

In this section, we present a new one-point zeroth-order momentum method that combines sample reuse as in Algorithm 2 in [14] and residual feedback as in Algorithm 3 in [14], but unlike Algorithm 2 in [14] that reuses samples only from the last iteration, it uses all past samples to update the agents' actions. The new algorithm is given as Algorithm 1.

Using the sampling strategy in (1), each agent plays the perturbed action $\hat{x}_{i,t}$ for n_t times and obtains n_t samples at

Algorithm 1 Risk-averse learning with momentum

Require: Initial value x_0, step size η, parameters a, b, δ, T, risk level α_i, $i = 1, \cdots, N$.

for episode $t = 1, \ldots, T$ **do**
  Select $n_t = \lceil bU^2(T-t+1)^a \rceil$
  Each agent samples $u_{i,t} \in \mathbb{S}^{d_i}$, $i = 1, \ldots, N$
  Each agent play $\hat{x}_{i,t} = x_{i,t} + \delta u_{i,t}$, $i = 1, \ldots, N$
  for $j = 1, \ldots, n_t$ **do**
    Let all agents play $\hat{x}_{i,t}$
    Obtain $J_i(\hat{x}_{i,t}, \hat{x}_{-i,t}, \xi_i^j)$
  end for
  for agent $i = 1, \ldots, N$ **do**
    Build EDF $\bar{F}_{i,t}(y)$
    Estimate CVaR: $\mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t}]$
    Construct gradient estimate $\bar{g}_{i,t} = \frac{d_i}{\delta}\left(\mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t-1}]\right) u_{i,t}$
    Update x: $x_{i,t+1} \leftarrow \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t} - \eta \bar{g}_{i,t})$
  end for
end for

episode t. For agent i, we denote the CDF of the random cost $J_i(\hat{x}_t, \xi_i)$ that is returned by the perturbed action $\hat{x}_t$ as $F_{i,t}^{\hat{x}_t}(y) = \mathbb{P}\{J_i(\hat{x}_t, \xi_i) \leq y\}$. Since $\hat{x}_t$ depends on t, we write $F_{i,t}$ for $F_{i,t}^{\hat{x}_t}$ for ease of notation. Using bandit feedback in the form of finitely many cost evaluations, the agents cannot obtain the accurate CDF $F_{i,t}$. Instead, they construct an empirical distribution function (EDF) $\hat{F}_{i,t}$ of the cost $J_i(\hat{x}_t, \xi_i)$ using n_t cost evaluations by

$$\hat{F}_{i,t}(y) = \frac{1}{n_t} \sum_{j=1}^{n_t} \mathbf{1}\{J_i(\hat{x}_t, \xi_i^j) \leq y\}. \tag{2}$$

To improve estimate of the CDF, we utilize past samples for all episodes $t \geq 2$, and construct a modified distribution estimate $\bar{F}_{i,t}$ by adding a momentum term:

$$\bar{F}_{i,t}(y) = \beta \bar{F}_{i,t-1}(y) + (1-\beta)\hat{F}_{i,t}(y), \tag{3}$$

where $\beta \in [0,1)$ is the momentum parameter. For $t = 1$, we set $\bar{F}_{i,t}(y) = \hat{F}_{i,t}(y)$. The agents utilize the distribution estimate $\bar{F}_{i,t}$ to calculate the CVaR estimates and further construct the gradient estimate using residual feedback as

$$\bar{g}_{i,t} = \frac{d_i}{\delta}\left(\mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t-1}]\right) u_{i,t}, \tag{4}$$

where δ is the size of the perturbation on the action $x_{i,t}$ defined above. Depending on the size of the perturbation, the played action $\hat{x}_{i,t}$ may be infeasible, i.e., $\hat{x}_{i,t} \notin \mathcal{X}_i$. To handle this issue, we define the projection set $\mathcal{X}_i^\delta = \{x_i \in \mathcal{X}_i | \mathrm{dist}(x_i, \partial\mathcal{X}_i) \geq \delta\}$ that we use in the projected gradient-descent update

$$x_{i,t+1} = \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t} - \eta \bar{g}_{i,t}). \tag{5}$$

Note that the agents are not able to obtain accurate values of $C_i(\hat{x}_t)$ using finite samples. In fact, if we use the distribution estimate $\bar{F}_{i,t}$ in (3) to calculate the CVaR values, there will

be a CVaR estimation error, which is defined as

$$\bar{\varepsilon}_{i,t} := \mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t}].$$

As a result, the gradient estimate $\bar{g}_{i,t}$ in (4) is biased since $\mathbb{E}[\bar{g}_{i,t}] = \mathbb{E}\left[\frac{d_i}{\delta}(C_i(\hat{x}_t) + \bar{\varepsilon}_{i,t})u_{i,t}\right] = \nabla_i C_i^{\delta}(x_t) + \mathbb{E}\left[\frac{d_i}{\delta}\bar{\varepsilon}_{i,t}u_{i,t}\right]$, with the bias captured by the last term. In what follows, we present a lemma that bounds the sum of the CVaR estimation errors.

Lemma 4. *Suppose that Assumption 3 holds. Then, the sum of the CVaR estimation errors satisfies*

$$\sum_{t=1}^{T} \|\bar{\varepsilon}_{i,t}\| \leq \frac{U}{\alpha_i}\Big(\frac{\beta}{1-\beta}e_{i,1} + \frac{\beta\Sigma_1}{1-\beta}T + \sum_{t=1}^{T} r_t\Big) \quad (6)$$

with probability at least $1-\gamma$, where $e_{i,t} := \sup_y |\bar{F}_{i,t}(y) - F_{i,t}(y)|$, $r_t := \sqrt{\frac{\ln(2T/\gamma)}{2bU^2(T-t+1)^a}}$, and $\Sigma_1 := C_0\eta\frac{d_i}{\delta}U\sqrt{N} + 2C_0\delta$.

Proof. To bound the CVaR estimation error, it suffices to bound the CDF differences using Lemma 3. Recall the definition of $\bar{F}_{i,t}$ in (3). Then, we have that

$$\begin{aligned}
&\sup_y |\bar{F}_{i,t}(y) - F_{i,t}(y)| \\
=&\sup_y |\beta\bar{F}_{i,t-1}(y) + (1-\beta)\hat{F}_{i,t}(y) - F_{i,t}(y)| \\
=&\sup_y |\beta(\bar{F}_{i,t-1}(y) - F_{i,t-1}(y)) + \beta(F_{i,t-1}(y) - F_{i,t}(y)) \\
&+ (1-\beta)(\hat{F}_{i,t}(y) - F_{i,t}(y))| \\
\leq&\beta\sup_y |\bar{F}_{i,t-1}(y) - F_{i,t-1}(y)| + \beta\sup_y |F_{i,t-1}(y) - F_{i,t}(y)| \\
&+ (1-\beta)\sup_y |\hat{F}_{i,t}(y) - F_{i,t}(y)|. \quad (7)
\end{aligned}$$

We now bound the latter two terms in the right-hand-side of (7) separately. Using Assumption 3 and Lemma 2, we have that

$$\begin{aligned}
&\sup_y |F_{i,t}(y) - F_{i,t-1}(y)| \\
\leq&C_0\|\hat{x}_t - \hat{x}_{t-1}\| \leq C_0\|x_t - x_{t-1}\| + 2C_0\delta \\
\leq&C_0\eta\|\bar{g}_t\| + 2C_0\delta \leq C_0\eta\frac{d_i}{\delta}U\sqrt{N} + 2C_0\delta := \Sigma_1, \quad (8)
\end{aligned}$$

where the last inequality holds since $\|\bar{g}_t\| = \sqrt{\sum_{i=1}^{N}\|\bar{g}_{i,t}\|^2} \leq \frac{\sqrt{N}d_iU}{\delta}$. Then, applying the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, we obtain that

$$\mathbb{P}\left\{\sup_y |F_{i,t}(y) - \hat{F}_{i,t}(y)| \geq \sqrt{\frac{\ln(2/\bar{\gamma})}{2n_t}}\right\} \leq \bar{\gamma}. \quad (9)$$

Define the events in (9) as A_t and denote $\gamma = \bar{\gamma}T$. Then, the following inequality holds

$$\sup_y |F_{i,t}(y) - \hat{F}_{i,t}(y)| \leq \sqrt{\frac{\ln(2T/\gamma)}{2n_t}}, \forall t = 1,\ldots,T, \quad (10)$$

with probability at least $1-\gamma$ since $1 - \mathbb{P}\{\bigcup_{t=1}^{T} A_t\} \geq 1 - \sum_{t=1}^{T}\mathbb{P}\{A_t\} \geq 1 - T\frac{\gamma}{T} \geq 1-\gamma$. Substituting the bounds in (8) and (10) into (7), and using the definition of $e_{i,t}$, we get that

$$e_{i,t} \leq \beta e_{i,t-1} + \beta\Sigma_1 + (1-\beta)r_t. \quad (11)$$

Note that (11) holds for all $t \in \{2,\cdots,T\}$. By iteratively using this inequality, we have that

$$e_{i,t} \leq \frac{\beta}{1-\beta}\Sigma_1 + (1-\beta)\sum_{k=0}^{k-2}\beta^k r_{t-k} + \beta^{t-1}e_{1,i}. \quad (12)$$

Taking the sum over $t = 1,\ldots,T$ of both sides of (12), we have that

$$\sum_{t=1}^{T} e_{i,t} \leq \frac{\beta}{1-\beta}e_{i,1} + \frac{\beta\Sigma_1}{1-\beta}T + \sum_{t=1}^{T} r_t, \quad (13)$$

with probability at least $1-\gamma$. Finally, using Lemma 3, we can obtain the desired result. □

Lemma 4 bounds the CVaR estimation error caused by using past samples. The next lemma quantifies the error due to residual feedback in the gradient estimate (4). Specifically, it bounds the second moment of the gradient estimate.

Lemma 5. *Let Assumptions 1, 2 and 3 hold. Suppose that the action is updated as in* (5)*. Then, the second moment of the gradient estimate satisfies*

$$\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq \frac{1}{1-\sigma}\|\bar{g}_1\|^2 + \frac{1}{1-\sigma}\Sigma_2 T, \quad (14)$$

where $\bar{g}_t := (\bar{g}_{i,t}, \bar{g}_{-i,t})$ is the concatenated vector of all agents' gradient estimates, $\sigma = \frac{4d_i^2L_0^2N\eta^2}{\delta^2}$, $\Sigma_2 = 16d_i^2N^2L_0^2 + \frac{48C_0^2d_i^4U^4N\eta^2\beta^2}{\delta^4(1-\beta)^2}(\sum_i\frac{1}{\alpha_i^2}) + \frac{192C_0^2d_i^2U^2\beta^2}{(1-\beta)^2}(\sum_i\frac{1}{\alpha_i^2}) + \frac{12\ln(2T/\gamma)d_i^2}{b\delta^2}(\sum_i\frac{1}{\alpha_i^2}) + \frac{24d_i^2U^2\beta^2e_m^2}{\delta^2}(\sum_i\frac{1}{\alpha_i^2})$, and $e_m := \max_i\{e_{i,1}\}$.

Proof. Using the fact that $r_t \leq \sqrt{\frac{\ln(2T/\gamma)}{2bU^2}} := r$ and (12), we have that

$$\begin{aligned}
\|e_{i,t}\|^2 &\leq \left\|\frac{\beta}{1-\beta}\Sigma_1 + r + \beta e_m\right\|^2 \\
&\leq 3\left\|\frac{\beta}{1-\beta}\Sigma_1\right\|^2 + 3\|r\|^2 + 3\|\beta e_m\|^2, \quad (15)
\end{aligned}$$

where the last inequality holds since $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$. Then, applying Lemma 3 to (15), we obtain that

$$\|\bar{\varepsilon}_{i,t}\|^2 \leq \frac{U^2}{\alpha_i^2}\left(\frac{3\beta^2\Sigma_1^2}{(1-\beta)^2} + 3\|r\|^2 + 3\beta^2\|e_m\|^2\right). \quad (16)$$

By the definition of $\bar{g}_{i,t}$ in (4), we have that

$$\begin{aligned}
\|\bar{g}_{i,t}\|^2 &= \frac{d_i^2}{\delta^2}\left((\mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t-1}])u_{i,t}\right)^2 \\
&\leq \frac{d_i^2}{\delta^2}\left(\mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t}] - \mathrm{CVaR}_{\alpha_i}[\bar{F}_{i,t-1}]\right)^2 \|u_{i,t}\|^2 \\
&\leq \frac{d_i^2}{\delta^2}\Big(2(\mathrm{CVaR}_{\alpha_i}[F_{i,t}] - \mathrm{CVaR}_{\alpha_i}[F_{i,t-1}])^2 \\
&\quad + 2(\bar{\varepsilon}_{i,t} - \bar{\varepsilon}_{i,t-1})^2\Big)\|u_{i,t}\|^2 \\
&\leq \frac{d_i^2}{\delta^2}\Big(2L_0^2\|\hat{x}_t - \hat{x}_{t-1}\|^2 + 4|\bar{\varepsilon}_{i,t}|^2 + 4|\bar{\varepsilon}_{i,t-1}|^2\Big), \qquad (17)
\end{aligned}$$

where the last inequality is due to the fact that the CVaR function is Lipschitz continuous as shown in Lemma 1. We now bound the term $\|\hat{x}_t - \hat{x}_{t-1}\|^2$ in (17). Recalling the update rule $x_{i,t} = \mathcal{P}_{\mathcal{X}_i^\delta}(x_{i,t-1} - \eta_i \bar{g}_{i,t-1})$, we have that

$$\begin{aligned}
\|\hat{x}_t - \hat{x}_{t-1}\|^2 &= \|x_t - x_{t-1} + \delta u_t - \delta u_{t-1}\|^2 \\
&\leq 2\|x_t - x_{t-1}\|^2 + 2\delta^2(2\|u_t\|^2 + 2\|u_{t-1}\|^2) \\
&\leq 2\|x_t - x_{t-1}\|^2 + 8N\delta^2 \leq 2\eta^2\|\bar{g}_{t-1}\|^2 + 8N\delta^2.
\end{aligned}$$

Substituting the above inequality and the inequality in (16) into the right hand side of (17) and summing both sides over all agents $i = 1, \ldots, N$, we have that

$$\begin{aligned}
\|\bar{g}_t\|^2 &= \sum_{i=1}^N \|\bar{g}_{i,t}\|^2 \\
&\leq \frac{d_i^2 N}{\delta^2}\Big(4L_0^2\eta^2\|\bar{g}_{t-1}\|^2 + 16L_0^2 N\delta^2\Big) \\
&\quad + \frac{8d_i^2}{\delta^2}\sum_{i=1}^N \frac{U^2}{\alpha_i^2}\left(\frac{3\beta^2\Sigma_1^2}{(1-\beta)^2} + 3\|r\|^2 + 3\beta^2\|e_m\|^2\right) \\
&\leq \sigma\|\bar{g}_{t-1}\|^2 + 16d_i^2N^2L_0^2 + \frac{12\ln(2T/\gamma)d_i^2}{b\delta^2}(\sum_i \frac{1}{\alpha_i^2}) \\
&\quad + \frac{24d_i^2U^2\beta^2}{\delta^2(1-\beta)^2}(C_0\eta\frac{d_i}{\delta}U\sqrt{N} + 2C_0\delta)^2(\sum_i \frac{1}{\alpha_i^2}) \\
&\quad + \frac{24d_i^2U^2\beta^2e_m^2}{\delta^2}(\sum_i \frac{1}{\alpha_i^2}) \\
&\leq \sigma\|\bar{g}_{t-1}\|^2 + \Sigma_2, \qquad (18)
\end{aligned}$$

where the last inequality is due to the fact that $(C_0\eta\frac{d_i}{\delta}U\sqrt{N} + 2C_0\delta)^2 \leq 2C_0^2\eta^2\frac{d_i^2}{\delta^2}U^2N + 8C_0^2\delta^2$. Summing (18) over all episode $t = 1, \ldots, T$ completes the proof. □

Before proving the main theorem, we present a regret decomposition lemma that links the regret to the errors bounds on the CVaR estimates and gradient estimates as provided in Lemmas 4 and 5.

Lemma 6. *Let Assumptions 1, 2 and 3 hold. Then, the regret of Algorithm 1 satisfies*

$$\begin{aligned}
\mathrm{R}_{C_i}(T) \leq &\frac{D_x^2}{2\eta} + \frac{\eta}{2}\mathbb{E}\big[\sum_{t=1}^T \|\bar{g}_{i,t}\|^2\big] + (4\sqrt{N} + \Omega)L_0\delta T \\
&+ \frac{d_i D_x}{\delta}\sum_{t=1}^T |\bar{\varepsilon}_{i,t}|, \qquad (19)
\end{aligned}$$

where $\Omega > 0$ is a constant that represents the error from projection $\mathcal{P}$.

Proof. The proof can be adapted from Lemma 5 in [14] and is omitted due to the space limitations. □

Theorem 1. *Let Assumptions 1, 2 and 3 hold and select $\eta = \frac{D_x}{d_i L_0 N}T^{-\frac{3a}{4}}$, $\delta = \frac{D_x}{N^{\frac{1}{6}}}T^{-\frac{a}{4}}$, $\beta = \frac{1}{U^2T^{\frac{a}{4}}}$. Suppose that n_t is chosen as in (1) with $a \in (0,1)$, and the EDF and the gradient estimate are defined as in (2) and (4), respectively. Then, when $T \geq (8N^{\frac{2}{3}})^{\frac{1}{a}}$, Algorithm 1 achieves regret*

$$\mathrm{R}_{C_i}(T) = \mathcal{O}(D_x d_i L_0 N S(\alpha)\ln(T/\gamma)T^{1-\frac{a}{4}}),$$

with probability at least $1-\gamma$ and where $S(\alpha) := \sum_{i=1}^N \frac{1}{\alpha_i^2}$.

Proof. Substituting the bounds in Lemmas 4 and 5 into Lemma 6, we have that

$$\begin{aligned}
\mathrm{R}_{C_i}(T) \leq &\frac{D_x^2}{2\eta} + \frac{\eta}{2}\left(\frac{1}{1-\sigma}\|\bar{g}_1\|^2 + \frac{1}{1-\sigma}\Sigma_2 T\right) \\
&+ \frac{d_i D_x U}{\delta\alpha_i}\Big(\frac{\beta}{1-\beta}e_{i,1} + \frac{\beta\Sigma_1}{1-\beta}T + \sum_{t=1}^T r_t\Big) \\
&+ (4\sqrt{N} + \Omega)L_0\delta T. \qquad (20)
\end{aligned}$$

Substituting δ, η and β into the above inequality, we can obtain the desired result. The detailed proof is straightforward and is omitted due to the space limitations. □

Note that Algorithm 1 in fact achieves the regret $\mathrm{R}_{C_i}(T) = \tilde{\mathcal{O}}(T^{1-\frac{a}{4}})$, in which we use the notation $\tilde{\mathcal{O}}$ to hide constant factors and poly-logarithmic factors of T. This result matches the best result in [14]. Compared to [14], the analysis of Algorithm 1 that combines historical information and residual feedback is nontrivial and requires an appropriate selection of the momentum parameter to show convergence. Note that Algorithm 3 in [14] is a special case of Algorithm 1 for the momentum parameter $\beta = 0$.

V. Numerical Experiments

In this section, we illustrate the proposed algorithm on a Cournot game. Specifically, we consider two risk-averse agents $i = 1, 2$ that compete with each other. Each agent determines the production level x_i and receives an individual cost feedback which is given by $J_i = 1 - (2 - \sum_j x_j)x_i + 0.2x_i + \xi_i x_i$, where $\xi_i \sim U(0,1)$ is a uniform random variable. Here we utilize the term $\xi_i x_i$ to represent the uncertainty occurred in the market, which is proportional to the production level x_i. The agents have their own risk levels α_i and they aim to minimize the risk of incurring high costs, i.e., the CVaR of their cost functions.

Note that the regret for agent i depends on the sequence of the other agent's actions $\{x_{-i,t}\}_{t=1}^T$, and this sequence depends on the algorithm. Therefore, it is not appropriate to compare different algorithms in terms of the regret. Instead, we use the empirical performance to evaluate the algorithms. Specifically, we compute the CVaR values at each episode, and observe how fast the agents can minimize the CVaR values during the learning process.

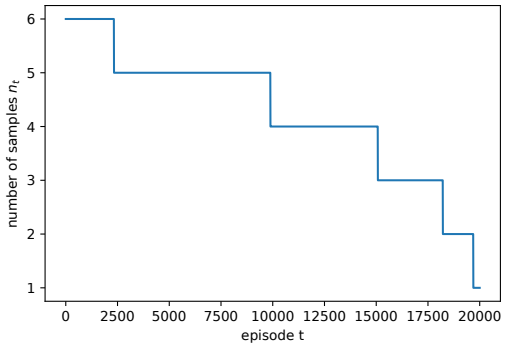

Fig. 1. The number of samples of Algorithm 1.

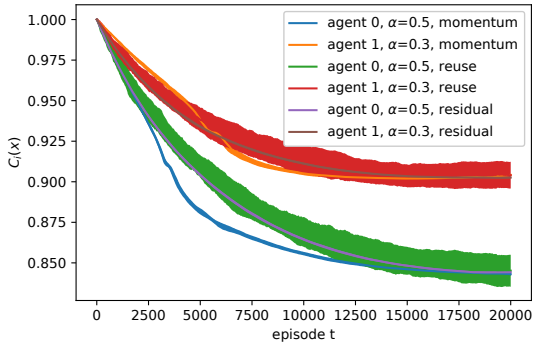

Fig. 2. Comparative results among our proposed zeroth-order momentum method, and the three algorithms in [14]. Shaded areas represent $\pm$ one standard deviation over 20 runs.

To compute the distribution function $\bar{F}_{i,t}$, we divide the interval $[0, U]$ into n bins of equal width, and we approximate the expectation by the sum of finite terms. We compare our zeroth-order momentum method with the two algorithms in [14], which we term as the algorithm with sample reuse and the algorithm with residual feedback, respectively. The number of samples n_t at each episode for Algorithm 1 is shown in Figure 1. We set the momentum parameter $\beta = 0.5$. Each algorithm is run for 20 trials and the parameters of these algorithms are separately optimally tuned. The CVaR values of these algorithms are presented in Figure 2. We observe that all the algorithms finally converge to the same CVaR values, but the zeroth-order momentum method proposed here converges faster. This is because our method uses all past samples to get a better estimate of the CVaR values and thus allows for a larger step size. Moreover, we also observe that the standard deviation of both our proposed algorithm and the algorithm with residual feedback is sufficiently small, which verifies the variance reduction effect of using residual feedback.

Moreover, we explore the effect of different momentum parameters β. In Figure 3, we present the achieved CVaR values of our method for various β values and all other parameters unchanged. Recalling the definition of the distribution estimate in (3), large values of β place more weight on

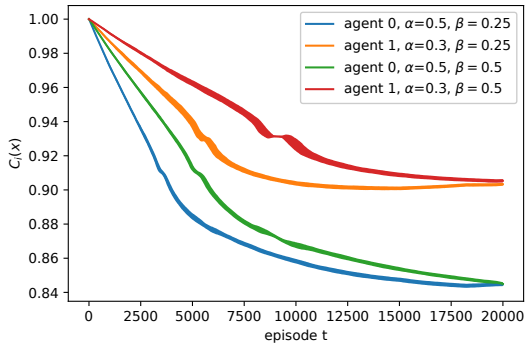

Fig. 3. CVaR values achieved by Algorithm 1 with different β values. Shaded areas represent $\pm$ one standard deviation over 20 runs.

outdated cost values and thus reduce the convergence speed.

VI. Conclusion

In this work, we proposed a zeroth-order momentum method for online convex games with risk-averse agents. The use of momentum that employs past samples allowed to improve the ability of the algorithm to estimate the CVaR values. Zeroth-order estimation of the CVaR gradients using residual feedback allowed us to reduce the variance of the CVaR gradient estimates. We showed that the proposed algorithm theoretically achieves no-regret learning with high probability, matching the best known result in literature. Moreover, we provided numerical simulations that demonstrated the superior performance of our method in practice. While sample reuse and residual feedback had been both separately shown to improve the performance of online learning, here we showed that their combination yields yet better performance.

References

[1] S. Shalev-Shwartz and Y. Singer, "Convex repeated games and fenchel duality," in *NIPS*, vol. 6. Citeseer, 2006, pp. 1265–1272.

[2] G. J. Gordon, A. Greenwald, and C. Marks, "No-regret learning in convex games," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 360–367.

[3] P. G. Sessa, I. Bogunovic, M. Kamgarpour, and A. Krause, "No-regret learning in unknown games with correlated payoffs," *Advances in Neural Information Processing Systems*, vol. 32, pp. 13 624–13 633, 2019.

[4] Y. Shi and B. Zhang, "No-regret learning in cournot games," *arXiv preprint arXiv:1906.06612*, 2019.

[5] E. Hazan, "Introduction to online convex optimization," *arXiv preprint arXiv:1909.05207*, 2019.

[6] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[7] W. F. Sharpe, "The sharpe ratio," *Journal of portfolio management*, vol. 21, no. 1, pp. 49–58, 1994.

[8] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical finance*, vol. 9, no. 3, pp. 203–228, 1999.

[9] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[10] T. Soma and Y. Yoshida, "Statistical learning with conditional value at risk," *arXiv preprint arXiv:2002.05826*, 2020.

[11] A. R. Cardoso and H. Xu, "Risk-averse stochastic convex bandit," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 39–47.

[12] S. Curi, K. Levy, S. Jegelka, A. Krause *et al.*, "Adaptive sampling for stochastic risk-averse learning," *arXiv preprint arXiv:1910.12511*, 2019.

[13] A. Tamkin, R. Keramati, C. Dann, and E. Brunskill, "Distributionally-aware exploration for cvar bandits," in *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.
[14] Z. Wang, Y. Shen, and M. Zavlanos, "Risk-averse no-regret learning in online convex games," in *International Conference on Machine Learning*. PMLR, 2022, pp. 22 999–23 017.
[15] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.
[16] Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "Boosting one-point derivative-free online optimization via residual feedback," *arXiv preprint arXiv:2010.07378*, 2020.
[17] M. Bravo, D. Leslie, and P. Mertikopoulos, "Bandit learning in concave n-person games," *Advances in Neural Information Processing Systems*, vol. 31, 2018.