Provable Optimization of Quantal Response Leader-Follower Games with Exponentially Large Action Spaces

Jinzhao Li
Purdue University
West Lafayette, IN, USA
li4255@purdue.edu

Daniel Fink
Cornell Lab of Ornithology
Ithaca, NY, USA
df36@cornell.edu

Christopher Wood
Cornell Lab of Ornithology
Ithaca, NY, USA
clw37@cornell.edu

Carla P. Gomes
Cornell University
Ithaca, NY, USA
gomes@cs.cornell.edu

Yexiang Xue
Purdue University
West Lafayette, IN, USA
yexiang@purdue.edu

ABSTRACT

Leader-follower games involve a leader committing strategies before her followers. We consider quantal response leader-follower games, where the followers' response is probabilistic due to their bounded rationality. Moreover, both the leader's and followers' action spaces are exponentially large with respect to the problem size, hence rendering the overall complexity to solve these games beyond NP-complete. We propose the XOR-Game algorithm, which converges in linear speed towards the equilibrium of convex quantal response leader-follower games (#P-hard to find the equilibrium even though convex). XOR-Game combines stochastic gradient descent with XOR-sampling, a provable sampling approach which transforms highly intractable probabilistic inference into queries to NP oracles. We tested XOR-Game on zero-sum and distribution matching leader-follower games. Experiments show XOR-Game converges faster to a good leader's strategy compared to several baselines. In particular, XOR-Game helps to find the optimal reward allocations for the Avicaching game in the citizen science domain, which harnesses rewards to motivate bird watchers towards tasks of high scientific value.

KEYWORDS

Leader-follower games; Quantal response; XOR-Game

ACM Reference Format:
Jinzhao Li, Daniel Fink, Christopher Wood, Carla P. Gomes, and Yexiang Xue. 2023. Provable Optimization of Quantal Response Leader-Follower Games with Exponentially Large Action Spaces. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 10 pages.

1 INTRODUCTION

Leader-follower games, also known as the Stackelberg games [13], involve leaders committing strategies before her followers. Over the years, leader-follower games have been studied extensively with their wide applications in security [44, 53], crowdsourcing [54, 55], AI for social good [19, 39], etc. Recent studies have focused on the participants' *bounded rationality* [43]. In other words, players do not always play the best moves due to imperfect information or limited computational capacity.

In quantal response leader-follower games, the followers take probabilistic actions due to their bounded rationality. In a Logistic quantal response game, a follower maximizes her utility, albeit together with extreme value distributed latent factors. In the eyes of an observer who does not know the latent factors, the follower's behavior is in an exponential family distribution. Quantal response games have been studied in security games [11] , and in games for social good [36]. The authors of [31] proposes an iterative approach to compute a near-optimal strategy for the leader in response to quantal responding adversaries.

In this paper, we consider solving quantal response leader-follower games, in which the sizes of the action spaces of both the leader and the followers grow exponentially quickly w.r.t. the problem size. Exponentially large action spaces are prevalent in real-world games, e.g., in real-time strategy (RTS) games [6] or security games [11]. They pose significant challenges in finding the equilibrium of the game because they prevents the leader from enumerating the entire action space, letting along reasoning about the followers' responses. Although several algorithms have been proposed in computing the equilibrium for normal form games [33, 40] and extensive form games [8, 48, 51], their mathematical programs involve summing over all the followers' actions. This becomes intractable as the number of actions grows exponentially in the size of the game.

We propose XOR-Game, the first algorithm which ***converges in linear speed towards the equilibrium of a convex quantal response leader-follower game with exponentially large action spaces***. Despite the game is convex with respect to the leader's strategy, the problem is still at least #P-hard due to the inference of the followers' actions from exponentially many probabilistic choices. Overall XOR-Game optimizes for the leader's objective following a Stochastic Gradient Descent (SGD) process. Our innovation is to harness XOR-sampling in the estimation of the gradient direction of each SGD step towards the optimal leader's strategy. XOR-sampling transforms the highly intractable (#P complete) probabilistic inference and sampling problems into queries to NP oracles using randomly generated XOR constraints. In another view, our XOR-Game algorithm transforms the highly intractable problem of reasoning about the followers' probabilistic actions into problems within the NP complexity class while obtaining provable guarantees on the linear convergence speed and the distances towards

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

the optimum. Notice other sampling approaches, e.g., MCMC sampling, provide unbiased samples only after an exponential number of burn-in steps. This is impossible in practice, and hence using these sampling approaches cannot result in similar convergence bounds as ours. Our guarantee is also significantly stronger than those offered by e.g., variational approaches [4, 24, 26–28, 42, 52], which are typically lower bounded only and can be arbitrarily loose. Even though the idea of incorporating XOR-sampling appears straightforward, all the theoretic derivation towards linear convergence guarantees cannot borrow from existing theoretic results of SGD. The key difficulty is due to that XOR-sampling only provides constant approximation guarantees for the probability of drawing samples but cannot guarantee unbiased sample estimation, which unfortunately was needed by most prior analysis.

Among many real-world applications, we consider two special cases of convex quantal response leader-follower games. The first is a zero-sum game, where the leader is to minimize the expected utility of the followers. The second is a distribution-matching leader-follower game, which the leader harnesses rewards to move the distribution of the followers' actions towards a given distribution. Our games have applications in the citizen science domain, where the organizer harnesses rewards to motivate citizen scientists towards tasks with high scientific value. In particular, we apply our game in the recently deployed *Avicaching* game [54, 55] in the *eBird* citizen science framework, where the organizer encourages bird enthusiasts towards bird watching activities in remote and undersampled locations. The experiment evaluations on both synthetic games as well as on real-world Avicaching games show that our XOR-Game is able to produce better leader's strategies in fewer SGD iterations compared to competing approaches.

2 PRELIMINARIES

2.1 Quantal Response Model

Classic decision theory takes the rational agent assumption, in which agents make perfect choices to maximize their utilities. However, this assumption falls short in the explanation of probabilistic decision-making and occasional deviation from optimal choices. Random utility models were developed to capture the bounded rationality of human decisions [5, 50]. Since its inception, random utility models have been used extensively in modeling human decision-making, ranging from demand prediction [3, 5, 50], behavior modeling [23, 25] to crowd-sourcing [17, 45].

Quantal response games, originated from Quantal choice model [32], were developed from random utility behavior models, and have achieved promising results modeling the bounded rationality of human beings [12, 37]. When faced with $N = 2^n$ choices, where the i-th choice has an observable utility value V_i, quantal response model assumes that the agent's choice a is to maximize the sum of the utility V_i and a latent factor ϵ_i:

$$a = \arg\max_{i \in \{1,\ldots,N=2^n\}} V_i + \epsilon_i. \tag{1}$$

ϵ_i is i.i.d. distributed in the standard Gumbel extreme value distribution Gumbel$(0, 1)$. In other words, the probabilistic nature of the agent's decision-making is due to the joint optimization of $V_i + \epsilon_i$ rather than V_i only. Noted that ϵ_i is only available to the agent but hidden from the observer. Gumbel noise is well accepted in literature to account for the stochasticity and/or irrationality of human decision-making [34, 45, 46]. Other types of noises, such as Gaussian noise, lead to other interesting models, such as the probit model. We leave as future work to consider those models. Finally, Gumbel(0,1) is used to simplify the theoretic derivation of our algorithm. Gumbel distributions with other parameters can be considered in a similar way.

Under the random utility model, it can be proven that in the eyes of an observer who do not have access to ϵ_i, the probability that the agent chooses the i-th option is given by the following exponential family distribution:

$$P(i) = \frac{\exp(V_i)}{Z} = \frac{\exp(V_i)}{\sum_{j=1}^{N=2^n} \exp(V_j)}. \tag{2}$$

Here, $Z = \sum_{j=1}^{N=2^n} \exp(V_j)$ is known as the partition function. Various quantities have been calculated for random utility behavior models, including the following expected utility:

Theorem 2.1. *[29] Under the random utility model, an agent's expected utility when following the decisions made based on Equation 1 is* $\log(\sum_{i=1}^{N=2^n} e^{V_i}) + \gamma$. *$\gamma$ is the Euler-Mascheroni constant.*

Exponential Action Spaces. In this paper, we consider games with exponential many choices. In other words, $N = O(2^n)$ and n denotes the input problem size. For example, in the Avicaching game in the citizen science domain where rewards are used to motivate crowdsourcing agents to explore sites with high scientific values, each agent's choice is represented as a set of locations that the agent explores. Suppose there are n locations, the set of choices are all the sets of locations, the size of which is 2^n.

2.2 Quantal Response Leader-Follower Games

Leader-follower games, also known as Stackelberg games, have attracted much research attention. In a game, the leader commits a strategy before her followers, often resulting in different equilibrium solutions from the Nash equilibrium where both sides commit strategies at the same time.

The random utility behavior model discussed in the previous section leads to quantal response leader-follower games. In a quantal response leader-follower game, the follower's decision-making follows a random utility model. The equilibrium of a quantal response leader-follower game can be computed through:

$$\begin{aligned} &\min_r L(\{V_i, P(i)\}, r), \\ &\text{s. t. } a = \arg\max_{i \in \{1,\ldots,N=2^n\}} V_i(r) + \epsilon_i, \\ &\qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0, 1),\ \forall i \in \{1, \ldots, N = 2^n\}. \end{aligned} \tag{3}$$

Here, the leader's objective is to minimize L. The follower's utility function for choice i has part V_i, observable by the leader and depends on the leader's strategy r. It also contains an extreme value distributed latent part ϵ_i, which is only available to the follower but hidden to the leader.

Compared with the standard leader-follower game, the the quantal response game brings in the probabilistic responses of the followers into consideration, which fits better to reality in many occasions. Due to the latent factor ϵ_i, the follower's action is probabilistic in the eye of the leader; namely, the follower takes the i-th action with probability $P(i)$, which has the exponential family form in Equation 2. The objective is written as $L(\{V_i, P(i)\}, r)$ showing that the leader's objective function can be dependent on his strategy r, the follower's observable utility V_i and/or the probabilities of making each decision $P(i)$. The formulation in Equation 3 can be used to model both cases where the leader takes pure or mixed strategies. In the pure strategy case, the leader's action r can be an indicator variable of which action to take. In the mixed strategy case, r becomes a vector listing the probability of taking each action. Since we assume the follower acts according to a quantal response model, she always plays probabilistic (hence mixed) strategies in the eyes of the leader.

2.2.1 Zero-sum Games. While the XOR-Game algorithm can provably optimize many quantal response leader-follower games, we mainly consider two variants for this paper. The first variant we consider is the zero sum case, where the leader is to minimize the expected utility of the follower. In other words, the leader's objective is (according to Theorem 2.1):

$$L_0(\{V_i, P(i)\}, r) = \log\left(\sum_{i=1}^{N=2^n} e^{V_i(r)}\right) + \gamma. \quad (4)$$

We consider a special case where V_i is linear in r, namely, $V_i(r) = \theta_i^T r + \phi_i$. Here θ_i measures the influence of rewards on the leader's action. ϕ_i represents the intrinsic utility in choosing action i and can vary across actions (even though it does not depend on r). We show that the game in this case is convex in r (proof is in the supplementary materials):

THEOREM 2.2. *When $V_i(r) = \theta_i^T r + \phi_i$, the zero-sum quantal response leader-follower game is convex in r. Moreover, the gradient $\nabla L_0(r)$ has the following form of an expectation:*

$$\nabla L_0(r) = \mathbb{E}_{P(i)}[\theta_i] = \sum_{i=1}^{N=2^n} P(i)\theta_i. \quad (5)$$

It is known in a zero-sum matrix game, the Stackelberg equilibrium matches exactly to the Nash equilibrium [49, 58]. Nevertheless, we would like to point out that in our definition of a quantal response game, the follower's action is designed to maximize her utility function V_i in addition to an unobserved ϵ_i (Equation 1), where V_i depends on the complete information of the leader's strategy r. This definition implicitly assumes the leader commits her strategy before the follower. The Nash equilibrium, in this setting, can be difficult to be properly defined.

2.2.2 Distribution Matching Games. Another variant we consider in this paper is the game where the leader would like to stimulate certain behaviors from the follower. In particular, the leader would like to match the probability distribution of the follower's actions P to a desired distribution Q. The difference between two distributions is measured by Kullback–Leibler (KL) divergence. In other words, the leader's objective is:

$$L_{DM}(\{V_i, P(i)\}, r) = KL(Q||P) = \sum_{i=1}^{N=2^n} Q(i)\log\left(\frac{Q(i)}{P(i)}\right). \quad (6)$$

This formulation is specially useful for mechanism design problems, e.g. in [20, 38, 41]. In cases where certain follower's actions increase the social welfare, the leader would set high Q values to promote these actions. Again, we focus on the special case where V_i is linear in r in this paper:

THEOREM 2.3. *When $V_i(r) = \theta_i^T r + \phi_i$ is linear in r, the distribution matching leader-follower game is convex in r and the gradient $\nabla L_{DM}(r)$ can be represented as:*

$$\nabla L_{DM} = \mathbb{E}_{P(i)}[\theta_i] - \mathbb{E}_{Q(i)}[\theta_i] = \sum_{i=1}^{N=2^n} P(i)\theta_i - \sum_{i=1}^{N=2^n} Q(i)\theta_i. \quad (7)$$

This theorem's proof is in the supplementary materials.

Avicaching Game in Citizen Science. As a specific example, we consider a leader-follower game in the citizen science domain, where the leader (citizen science game organizer) harnesses limited rewards to encourage citizen scientists (followers) to conduct observations in remote and undersampled locations. We look into the Avicaching game hosted in the eBird citizen science program [54], where rewards are used to encourage bird watchers (citizen scientists, or followers) to visit undersampled Avicaching sites, which have large scientific values but are inconvenient and/or less interesting to visit than traditional hotspots. Each bird watcher's choice is characterized by a set of locations $L \subseteq \{l_1, \ldots, l_n\}$, which represents the set of spots one plan to visit during a bird watching trip. The organizer, in this case, harnesses reward $r = (r_1, \ldots, r_n)$ to stimulate visits to under-sampled locations. Here, r_i is the reward that a bird watcher receives when he visits location l_i. The bird watchers' (followers') utility $V_L(r)$ models both the intrinsic utilities to visit the location set L as well as the reward received.

2.3 XOR Sampling

Sampling from a combinatorial space has a formal complexity of #P-complete, the difficulty of which is beyond NP-completeness. Luckily, the recently proposed XOR-Sampling algorithm, as the result of a rich line of works using streamlining randomized constraints [1, 2, 9, 15, 16, 21, 22], provides a constant approximation guarantee on the probabilities of the samples generated. XOR-sampling transforms the highly intractable sampling problem into queries to NP-oracles while obtaining provable guarantees.

The high-level idea of XOR-sampling is to harness randomized constraints to guarantee the randomness of the samples generated. Consider a simple case where $w(x) : \{0,1\}^n \rightarrow \{0,1\}$ is a binary function and one would like to obtain a sample from the solution space $\{x : w(x) = 1\}$ uniformly at random. Querying a NP oracle returns one x satisfying $w(x) = 1$ albeit not at random. XOR-sampling works by querying NP oracles to find x which satisfies $w(x) = 1$ and subject to a few randomized XOR constraints. It can be proven that each additional XOR constraint removes approximately half of the solutions to $w(x) = 1$ at random. Hence, once a desirable number of XOR constraints are added and the resulting space has only one solution, it can be proven that the only solution remaining is a random one from the original space $\{x : w(x) = 1\}$. In this way, XOR-sampling is able to bound the probability of obtaining each sample within a constant multiplicative factor of its ground-truth probability. For general weighted functions XOR-sampling has similar guarantees, although the sampling process becomes more complex. Our proposed XOR-Game algorithm depends on the following approximation bounds:

THEOREM 2.4. *(Ermon et al., 2013)*[1] *Let* $1 < \delta < \sqrt{2}$, $0 < \gamma < 1$, $w : \{0,1\}^n \rightarrow \mathbb{R}^+$ *be an unnormalized weight function.* $P(x) \propto w(x)$ *is the normalized distribution. Then, with probability at least* $1-\gamma$, *XOR-Sampling*(w, δ, γ) *succeeds and outputs a sample* x_0 *using* $O(-n \log(1-1/\sqrt{\delta}) \log(-n/\gamma \log(1-1/\sqrt{\delta})))$ *NP-oracle queries. Upon success, each* x_0 *is produced with probability* $P'(x_0)$. *We have*

$$\frac{1}{\delta} P(x_0) \leq P'(x_0) \leq \delta P(x_0).$$

Moreover, let $\phi : \{0,1\}^n \rightarrow \mathbb{R}$ *be a function mapping binary vectors to* $\mathbb{R}$. *Denote* $\phi(x)^+ = \max\{\phi(x), 0\}$ *and* $\phi(x)^- = \min\{\phi(x), 0\}$ *as the positive and negative part of* $\phi(x)$. *Then the expectation of one sampled* $\phi(x)$ *satisfies,*

$$\frac{1}{\delta} \mathbb{E}_{P(x)}[\phi(x)^+] \leq \mathbb{E}_{P'(x)}[\phi(x)^+] \leq \delta \mathbb{E}_{P(x)}[\phi(x)^+],$$

$$\delta \mathbb{E}_{P(x)}[\phi(x)^-] \leq \mathbb{E}_{P'(x)}[\phi(x)^-] \leq \frac{1}{\delta} \mathbb{E}_{P(x)}[\phi(x)^-].$$

Algorithm 1: XOR-Game$_0$

```
Input   : r_0, {θ_i}_{i=1}^N, {φ_i}_{i=1}^N
Params  : T, K, η, δ, γ
for t = 0 to T do
    k ← 1;
    while k ≤ K do
        P(i) ∝ exp(θ_i^T r_t + φ_i);
        l' ← XOR-Sampling(P(i), δ, γ);
        if l' ≠ Failure then
            l'_k ← l';
            k ← k + 1;
        end
    end
    ḡ_t ← (1/K) Σ_{k=1}^K θ_{l'_k};
    r_{t+1} = r_t − η ḡ_t;
end
return r̄_T = (1/T) Σ_{t=1}^T r_t
```

3 XOR-GAME

The challenge in solving quantal response leader follower games is the intractable probabilistic inference over the follower's strategies. In this paper, we consider games in which the follower's action space is exponentially large. In other words, the number of actions N is of size $O(2^n)$, where n is the problem size. These games are prevalent in real world. See the Avicaching game presented in the experiment section for an example. Notice we assume there are an compact representation for all θ_i's and ϕ_i's. Even though there are $2N$ vectors of these, we assume the availability of efficient functions $\theta(i)$ and $\phi(i)$. When given i, they return θ_i and ϕ_i, respectively. The length of encoding both functions $\theta(i)$ and $\phi(i)$ are within $O(n)$, i.e., the length of the problem description. In this setup, the quantal

[1]The details of the discretization scheme and the choices of parameters of the original algorithm which yield the bound of this form is discussed in [14].

Algorithm 2: XOR-Game$_{DM}$

```
Input   : r_0, {Q(i)}_{i=1}^N, {θ_i}_{i=1}^N, {φ_i}_{i=1}^N
Params  : T, K, S, η, δ, γ
j ← 1;
while j ≤ S do
    l'' ← XOR-Sampling(Q, δ, γ);
    if l'' ≠ Failure then
        l''_j ← l''; j ← j + 1;
    end
end
for t = 0 to T do
    k ← 1;
    while k ≤ K do
        P(i) ∝ exp(θ_i^T r_t + φ_i);
        l' ← XOR-Sampling(P(i), δ, γ);
        if l' ≠ Failure then
            l'_k ← l'; k ← k + 1;
        end
    end
    ḡ_t ← (1/K) Σ_{k=1}^K θ_{l'_k} − (1/S) Σ_{j=1}^S θ_{l''_j};
    r_{t+1} = r_t − η ḡ_t;
end
return r̄_T = (1/T) Σ_{t=1}^T r_t
```

response leader-follower game is at least #P-hard, even limiting to the convex games considered in Theorem 2.2 and 2.3. This is because it is already #P-hard to compute the partition function in $P(i)$ in Equation 2. In other words, it is already #P-complete to evaluate the leader's objective function even for a fixed strategy.

We propose XOR-Game, which ***converges towards the equilibrium of convex quantal response leader-follower games in linear number of stochastic gradient descent iterations***. The XOR-Game algorithm should enjoy the convergence bound for a wide variety of quantal response games. However, the actual algorithms and the convergence bounds slightly differ across different game setups, due to differences in estimating the derivatives and their correspondingly different approximation bounds given by XOR-sampling. In this paper, we demonstrate such convergence bounds on the aforementioned zero-sum and distribution matching quantal response leader-follower games. However, we are confident that similar guarantees generalize to many other games.

The algorithm variants for solving zero sum game and distribution matching game are shown in Algorithm 1 and Algorithm 2. The procedures of XOR-Game$_0$ and XOR-Game$_{DM}$ have minimal differences. Both algorithms apply SGD to find the optimal reward r that minimizes the leader's objective. r_0 is the initialization of the reward vector. The follower's observable utility is $V_i(r) = \theta_i^T r + \phi_i$. Samples generated from XOR-sampling are used to estimate the expectations in the gradient calculation (according to Equation 5 and 7). Because XOR-Sampling has a failure rate, repeated sampling is used until a desired number of samples are obtained. K samples are drawn from the behavior model of followers, and S

samples are from the targeting model. XOR-Sampling takes parameters (δ, γ). After the gradient estimation, r_{t+1} from the next iteration moves from r_t following the negative gradient direction. η is the step size of SGD. Finally after T SGD steps, the average of $r_1, \ldots, r_T$ is returned as the output. Denote the total variance $Var_{P(i)}(\theta_i) = \mathbb{E}_{P(i)}(||\theta_i||_2^2) - ||\mathbb{E}_{P(i)}(\theta_i)||_2^2$. We can show that the convergence bound for XOR-Game$_0$ to solve the zero sum game is:

Theorem 3.1. *(Convergence for zero-sum game) In a zero sum quantal response leader follower game with objective in Equation 4 and* $V_i(r) = \theta_i^T r + \phi_i$, r^* *attains the minimum of the leader's objective.* $\overline{r_T}$ *is the output of XOR-Game*$_0$ *starting from* r_0 *and running T SGD iterations. In iteration t of SGD,* g_t *is the estimated gradient, i.e.,* $r_{t+1} = r_t - \eta g_t$. *If* $\max_P Var_{P(i)}(\theta_i) \leq \sigma^2$, $||r_t - r^*||_2 \leq R$, $\eta \leq (2-\delta^2)/(\sigma^2\delta)$, $\max_P ||\mathbb{E}_{P(i)}(\theta_i^+)||_2 \leq G$, *and* $\max_P ||\mathbb{E}_{P(i)}(\theta_i^-)||_2 \leq G$, *where* $\theta^+ = \max\{\theta, 0\}$ *and* $\theta^- = \min\{\theta, 0\}$, *we have*

$$\mathbb{E}[L_0(\overline{r_T})] - L_0(r^*) \leq \frac{\delta||r_0 - r^*||_2^2}{2\eta T} + \frac{\eta}{K}(\sigma^2 + 2G^2) + 2\eta\delta(\delta^2 - 1)G^2 + 2(\delta^2 - 1)GR.$$

Theorem 3.1 proves that XOR-Game$_0$ converges to the equilibrium of the zero-sum game in a linear number of SGD steps in addition to a few vanishing terms. Here, the first term on the right-hand side scales inversely proportional to the number of SGD iterations T, suggesting a linear convergence speed towards the equilibrium. The second term can be reduced by increasing K, the number of XOR samples in estimating $\mathbb{E}_{P(i)}[\theta_i]$. The third and the fourth terms are the products of constants with $(\delta^2 - 1)$, which can be minimized with a more accurate, yet more time-consuming XOR-sampling (bringing δ closer to 1). The proof of Theorem 3.1 shares the same high-level idea with Theorem 3.4. The detailed proof is left to the supplementary materials. The convergence bound for XOR-Game$_{DM}$, the algorithm variant to solve the distribution matching game depends on a stronger condition:

Definition 3.2. (Match signs at every dimension) A group of vectors $\Theta = \{\theta_1, \ldots, \theta_N\}$ matches signs at every dimension, if for any two vectors $\theta_i, \theta_j \in \Theta$, $\theta_i = (\theta_{i1}, \ldots, \theta_{iL})^T$, $\theta_j = (\theta_{j1}, \ldots, \theta_{jL})^T$, for any dimension $k \in \{1, \ldots, L\}$, we have $\theta_{ik}\theta_{jk} \geq 0$.

The provable guarantee for XOR-Game$_{DM}$ requires all θ_i in the distribution matching game to match signs at every dimension. This requirement is not too stringent. As we have pointed out, distribution matching leader follower games are usually seen in mechanism design problems, where the leader searches for a strategy to maximize certain behaviors from the followers. Here, the leader's strategy r typically represents the incentives offered to the follower. θ_i in this case becomes indicator variables whether certain incentives are earned if the follower takes action i. Due to this reason, all θ_i are non-negative, satisfying the matching signs condition. With these definitions, the convergence bound for the distribution matching leader follower game is as follows:

Theorem 3.3. *(Convergence for distribution matching game) Suppose a distribution matching leader-follower game has the objective in Equation 6.* $V_i(r) = \theta_i^T r + \phi_i$. *Denote* r^* *as the optimal leader's strategy.* $\overline{r_T}$ *is the output of the XOR-Game*$_{DM}$. *Suppose* $\{\theta_1, \ldots, \theta_N\}$ *match signs at every dimension,* $\max_P Var_{P(i)}(\theta_i) \leq \sigma^2$, $\max_P ||\mathbb{E}_{P(i)}(\theta_i)||_2 \leq G$, $Var_{Q(i)}(\theta_i) \leq \sigma^2$, $||\mathbb{E}_{Q(i)}(\theta_i)||_2 \leq G$, *when* $1 < \delta < \sqrt{2}$ *is used in XOR-sampling and the SGD step size* $\eta \leq (2-\delta^2)/(\sigma^2\delta)$, $||r_i - r^*||_2 \leq R$ *for all* $r_1, \ldots, r_T$, *we have:*

$$\mathbb{E}[L_{DM}(\overline{r_T})] - L_{DM}(r^*) \leq \frac{\delta||r_0 - r^*||_2^2}{2\eta T} + \tag{8}$$

$$(\delta^2 - 1)\left[\sqrt{2}GR + 2\eta\left(\frac{\sigma^2 + G^2}{\min\{K, S\}} + \delta G^2\right)\right] + 2\eta(\delta^2 + 1)\frac{\sigma^2 + G^2}{\min\{K, S\}}.$$

To interpret this inequality, the first term on the right-hand side of inequality 8 scales inversely proportional to T, showing a linear convergence speed towards the optimal leader's strategy r^*. The second term is the product of $(\delta^2 - 1)$ and a constant (all terms in the square bracket). This term can be minimized with more accurate (yet more expensive) XOR-sampling, bringing in δ closer to 1. The term in the second line scales inversely proportional to $\min\{K, S\}$, which can be minimized by increasing K and S; e.g., drawing more samples. In summary, this theorem still shows a linear convergence bound and two tails terms which can be minimized via better sampling. The proof of Theorem 3.3 depends on the following Theorem 3.4, which was motivated by Theorem 3 in [14]. Nevertheless, Theorem 3 in [14] does not apply to the case where the difference of two XOR sampling processes are used to estimate the gradient. We therefore need to come up with novel proof techniques for distribution matching games, which yields the following Theorem 3.4:

Theorem 3.4. *Suppose function* $f : \mathbb{R}^d \to \mathbb{R}$ *is L-smooth convex.* $r^* = \arg\min_r f(r)$. *At any point r, the gradient* $\nabla f(r)$ *can be decomposed into* $\nabla f(r) = \nabla p(r) - \nabla q(r)$. *At the t-th iteration of SGD,* $g_t = k_t - l_t$ *is the estimated gradient, i.e.,* $r_{t+1} = r_t - \eta g_t$. k_t *(or* l_t*) is the estimation of* $\nabla p(r_t)$ *(or* $\nabla q(r_t)$*).* $\{k_t, l_t, \nabla p(r_t), \nabla q(r_t)\}$ *match signs at every dimension. If* $Var(k_t) \leq \sigma^2$, $Var(l_t) \leq \sigma^2$, $||\mathbb{E}[k_t]||_2^2 \leq G^2$, $||\mathbb{E}[l_t]||_2^2 \leq G^2$, *and there exists* $1 < c < \sqrt{2}$, *s.t.*

$$\frac{1}{c}[\nabla p(r_t)]^+ \leq \mathbb{E}[k_t^+] \leq c[\nabla p(r_t)]^+$$
$$c[\nabla p(r_t)]^- \leq \mathbb{E}[k_t^-] \leq \frac{1}{c}[\nabla p(r_t)]^-$$
$$\frac{1}{c}[\nabla q(r_t)]^+ \leq \mathbb{E}[l_t^+] \leq c[\nabla q(r_t)]^+$$
$$c[\nabla q(r_t)]^- \leq \mathbb{E}[l_t^-] \leq \frac{1}{c}[\nabla q(r_t)]^-.$$

Let $R = \max_t ||r_t - r^*||$, *with* $\eta \leq \frac{2-c^2}{Lc}$, $\overline{r_T} = \frac{1}{T}\sum_{t=1}^T r_t$, *we have:*

$$\mathbb{E}[f(\overline{r_T})] - f(r^*) \leq \tag{9}$$

$$\frac{c}{2\eta T}||r_0 - r^*||_2^2 + \left(c - \frac{1}{c}\right)\left(\sqrt{2}GR + 2\eta(\sigma^2 + G^2)\right) + 2\eta\left(c + \frac{1}{c}\right)\sigma^2.$$

The proof of Theorem 3.3 is to apply Theorem 3.4 to the objective function of XOR-Game$_{DM}$. Notice that L_{DM} is L-smooth when the total variation $\max_P Var_{P(i)}(\theta_i)$ is bounded (proved in a lemma). Those 4 constraints on the expectation of estimated gradients can be achieved by tuning parameters of XOR-Sampling.

To prove Theorem 3.4, we need the following lemmas. The proofs of these lemmas are left in the supplementary materials:

Lemma 3.5. *Suppose f is convex.* $r^* = \arg\min_r f(r)$. *At the t-th iteration of SGD,* $g_t = k_t - l_t$ *is the estimated gradient.* $\{k_t, l_t, \nabla p(r_t),$

$\nabla q(r_t)\}$ match signs at every dimension, and there exists $1 < c < \sqrt{2}$, s.t. $\frac{1}{c}[\nabla p(r_t)]^+ \le \mathbb{E}[k_t^+] \le c[\nabla p(r_t)]^+$, $c[\nabla p(r_t)]^- \le \mathbb{E}[k_t^-] \le \frac{1}{c}[\nabla p(r_t)]^-$, $\frac{1}{c}[\nabla q(r_t)]^+ \le \mathbb{E}[l_t^+] \le c[\nabla q(r_t)]^+$, $c[\nabla q(r_t)]^- \le \mathbb{E}[l_t^-] \le \frac{1}{c}[\nabla q(r_t)]^-$, we have:

$$\langle \nabla p(r_t), \mathbb{E}[k_t] \rangle \ge \frac{1}{c} ||\mathbb{E}[k_t]||_2^2, \tag{10}$$

$$\langle \nabla q(r_t), \mathbb{E}[l_t] \rangle \ge \frac{1}{c} ||\mathbb{E}[l_t]||_2^2, \tag{11}$$

$$\langle \nabla p(r_t), \mathbb{E}[l_t] \rangle \le c \langle \mathbb{E}[k_t], \mathbb{E}[l_t] \rangle, \tag{12}$$

$$\langle \nabla q(r_t), \mathbb{E}[k_t] \rangle \le c \langle \mathbb{E}[k_t], \mathbb{E}[l_t] \rangle. \tag{13}$$

Lemma 3.6. *Suppose all variables and pre-conditions are defined as in Theorem 3.4. In particular, $Var(k_t) \le \sigma^2$, $Var(l_t) \le \sigma^2$, $||\mathbb{E}[k_t]||_2^2 \le G^2$, $||\mathbb{E}[l_t]||_2^2 \le G^2$, we have*

$$||Tr[Cov(k_t, l_t)]|| = |\mathbb{E}[\langle k_t, l_t \rangle] - \langle \mathbb{E}[k_t], \mathbb{E}[l_t] \rangle| \le \sigma^2, \tag{14}$$

$$\mathbb{E}[\langle k_t, l_t \rangle] \le \sigma^2 + G^2. \tag{15}$$

Lemma 3.7. *Suppose all variables and conditions are defined as in Theorem 3.4. We have:*

$$\langle \nabla f(r_t), r_t - r^* \rangle \le c \langle \mathbb{E}[k_t] - \mathbb{E}[l_t], r_t - r^* \rangle + \sqrt{2} \left(c - \frac{1}{c} \right) GR. \tag{16}$$

Proof. (Formal proof of Theorem 3.4) By L-smoothness of f, for the t-th iteration,

$$\begin{aligned} f(r_{t+1}) &\le f(r_t) + \langle \nabla f(r_t), r_{t+1} - r_t \rangle + \frac{L}{2} ||r_{t+1} - r_t||_2^2 \\ &= f(r_t) - \eta \langle \nabla p(r_t) - \nabla q(r_t), k_t - l_t \rangle + \frac{L\eta^2}{2} ||k_t - l_t||^2 \\ &= f(r_t) + \frac{L\eta^2}{2} ||k_t - l_t||^2 - \\ &\quad \eta \left(\langle \nabla p(r_t), k_t \rangle - \langle \nabla q(r_t), k_t \rangle - \langle \nabla p(r_t), l_t \rangle + \langle \nabla q(r_t), l_t \rangle \right). \end{aligned}$$

Take the expectation w.r.t. k_t and l_t on both sides, and notice Equations 10, 11, 12, 13 in Lemma 3.5, we have:

$$\begin{aligned} \mathbb{E}[f(r_{t+1})] \le f(r_t) - \frac{\eta}{c} [||\mathbb{E}[k_t]||_2^2 + ||\mathbb{E}[l_t]||_2^2] + \\ 2\eta c \langle \mathbb{E}[k_t], \mathbb{E}[l_t] \rangle + \frac{L\eta^2}{2} \mathbb{E}[||k_t - l_t||^2]. \end{aligned}$$

Notice $Var(k_t) = \mathbb{E}[||k_t||_2^2] - ||\mathbb{E}[k_t]||_2^2$, $Var(l_t) = \mathbb{E}[||l_t||_2^2] - ||\mathbb{E}[l_t]||_2^2$, and $Tr[Cov(k_t, l_t)] = \mathbb{E}[\langle k_t, l_t \rangle] - \langle \mathbb{E}[k_t], \mathbb{E}[l_t] \rangle$, we further rewrite the right-hand side as:

$$\begin{aligned} &\mathbb{E}[f(r_{t+1})] \\ &\le f(r_t) - \frac{\eta}{c} \left[\mathbb{E}[||k_t||_2^2] - Var(k_t) + \mathbb{E}[||l_t||_2^2] - Var(l_t) \right] + \\ &2\eta c [\mathbb{E}[\langle k_t, l_t \rangle] - Tr[Cov(k_t, l_t)]] + \frac{L\eta^2}{2} \mathbb{E}[||k_t - l_t||^2]. \end{aligned}$$

After re-arranging terms, the right-hand side again becomes:

$$\mathbb{E}[f(r_{t+1})] \le f(r_t) - \frac{\eta(2 - L\eta c)}{2c} \mathbb{E}[||k_t - l_t||^2] + \text{tail}. \tag{17}$$

$$\begin{aligned} \text{tail} = &\frac{\eta}{c}(Var(k_t) + Var(l_t)) - 2\eta c Tr[Cov(k_t, l_t)] + \\ &2\eta(c - \frac{1}{c}) \mathbb{E}[\langle k_t, l_t \rangle]. \end{aligned}$$

Using $Var(k_t) \le \sigma^2$, $Var(l_t) \le \sigma^2$ and Lemma 3.6, we have:

$$\text{tail} \le 2\eta \left(c - \frac{1}{c} \right) (\sigma^2 + G^2) + 2\eta \left(c + \frac{1}{c} \right) \sigma^2. \tag{18}$$

For simplicity, denote the right-hand side of the previous inequality as a constant C_1. Hence, Equation 17 becomes:

$$\mathbb{E}[f(r_{t+1})] \le f(r_t) - \frac{\eta(2 - L\eta c)}{2c} \mathbb{E}[||k_t - l_t||^2] + C_1.$$

Using $\eta \le \frac{2-c^2}{Lc}$, we can further simplify this inequality to:

$$\mathbb{E}[f(r_{t+1})] \le f(r_t) - \frac{\eta c}{2} \mathbb{E}[||k_t - l_t||^2] + C_1.$$

Because f is convex, $f(r_t) \le f(r^*) + \langle \nabla f(r_t), r_t - r^* \rangle$. Follow the previous inequality we get:

$$\mathbb{E}[f(r_{t+1})] \le f(r^*) + \langle \nabla f(r_t), r_t - r^* \rangle - \frac{\eta c}{2} \mathbb{E}[||k_t - l_t||^2] + C_1.$$

Because of Lemma 3.7, we can further rewrite the previous inequality as

$$\begin{aligned} \mathbb{E}[f(r_{t+1})] \le f(r^*) + c \langle \mathbb{E}[k_t] - \mathbb{E}[l_t], r_t - r^* \rangle + \\ \sqrt{2} \left(c - \frac{1}{c} \right) GR - \frac{\eta c}{2} \mathbb{E}[||k_t - l_t||^2] + C_1. \end{aligned}$$

Define $C_2 = C_1 + \sqrt{2}\left(c - \frac{1}{c} \right) GR$, and notice $k_t - l_t = g_t$, we can write

$$\begin{aligned} \mathbb{E}[f(r_{t+1})] &\le f(r^*) + c \langle \mathbb{E}[g_t], r_t - r^* \rangle - \frac{\eta c}{2} \mathbb{E}[||g_t||^2] + C_2 \\ &= f(r^*) + \frac{c}{2\eta} \left(2\eta \langle \mathbb{E}[g_t], r_t - r^* \rangle - \eta^2 \mathbb{E}[||g_t||_2^2] \right) + C_2 \\ &= f(r^*) + \frac{c}{2\eta} \mathbb{E} \left[2\eta \langle g_t, r_t - r^* \rangle - \eta^2 ||g_t||_2^2 \right] + C_2. \end{aligned}$$

From the second last to the last equation, we also take expectation w.r.t. r_t on both sides. The equality holds because the randomness of g_t come from the sampling step at the t-th iteration, which is independent of r_t (whose randomness come from the first $t - 1$ iterations). Because $r_{t+1} = r_t - \eta g_t$, we have $||r_t - r^*||_2^2 - ||r_{t+1} - r^*||_2^2 = 2\eta \langle g_t, r_t - r^* \rangle - \eta^2 ||g_t||_2^2$. Hence we have:

$$\mathbb{E}[f(r_{t+1})] \le f(r^*) + \frac{c}{2\eta} \mathbb{E}[||r_t - r^*||_2^2 - ||r_{t+1} - r^*||_2^2] + C_2. \tag{19}$$

By summing up Equation 19 for $t = 0, \ldots, T - 1$, we get

$$\sum_{t=0}^{T-1} \mathbb{E}[f(r_{t+1}) - f(r^*)] \le \frac{c||r_0 - r^*||_2^2}{2\eta} + TC_2. \tag{20}$$

Finally, by Jensen's inequality, $Tf(\overline{r_T}) \le \sum_{t=1}^{T} f(r_t)$,

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[f(r_{t+1}) - f(r^*)] &= \mathbb{E}[\sum_{t=1}^{T} f(r_t)] - Tf(r^*) \\ &\ge T\mathbb{E}[f(\overline{r_T})] - Tf(r^*). \end{aligned} \tag{21}$$

Combining equations 20, 21, we have

$$\mathbb{E}[f(\overline{r_T})] \le f(r^*) + \frac{c||r_0 - r^*||_2^2}{2\eta T} + C_2, \tag{22}$$

which is exactly the equation in Theorem 3.4. □

To quantify the computational complexity of XOR-Game, we prove the following theorem in the supplementary materials detailing the number of queries to NP oracles needed for XOR-Game$_{DM}$ and XOR-Game$_0$. The proof of this Theorem is again left in the supplementary materials.

THEOREM 3.8. *XOR-Game$_0$ in Algorithm 1 uses $O(-Tn\log(1-1/\sqrt{\delta})\log(-n/\gamma\log(1-1/\sqrt{\delta}))+TK)$ queries to NP oracles. XOR-Game$_{DM}$ in Algorithm 2 uses $O(-Tn\log(1-1/\sqrt{\delta})\log(n/\gamma\log(1-1/\sqrt{\delta}))+TK+S)$ queries to NP oracles.*

4 EXPERIMENTS

We demonstrate empirical evidence that XOR-Game outperforms a few competing approaches in the speed and the quality of the solutions found for both the quantal response zero-sum leader-follower games and distribution-matching games. Our evaluation is conducted on a synthetic benchmark set and a behavior model learned from real-world data obtained from the Avicaching game, which promotes bird watchers to collect data in remote and undersampled locations using the so-called Avicaching points [54, 55]. The baseline approaches we consider are: (1) BRQR algorithm [56], which minimizes the leader's objective in quantal response stackelberg games. Their approach is based on a full gradient descend (GD) optimizer, hence needs to go over all the follower's actions in each iteration and is only applicable on games with small action spaces. (2) gibbs_game, which uses SGD to minimize the leader's objective but utilizes Gibbs sampling in the estimation of the gradient direction. (3) bp_game, which uses samples generated from the marginal probabilities computed via loopy belief propagation during SGD, [30, 35, 57] and (4) cbp_game, which harnesses the recently proposed BP chain method in generating samples in SGD [18]. For the fairness of comparisons, the leader's objective L_0 for the zero-sum game is computed using an exact model counter Ace [10]. The leader's objective L_{DM} for the distribution matching game is the KL-divergence, which is computed using Ace and XOR-sampling. The estimated KL-divergence is close to the groundtruth due to the constant approximation guarantee of XOR-sampling and the exactness of Ace. Additional details are in the supplementary materials.

In synthetic and real-world experiments, we use the Avicaching game as the background. In the Avicaching game, the leader (Avicaching game organizer) harnesses rewards to motivate the followers (bird watchers) to visit remote and under-sampled locations. The rewards are in the form of virtual Avicaching points, which marks the participants' contributions to science. At the end of each season of the Avicaching game, a lottery is drawn from which Avicaching participants have opportunities to win birding gears based on how many Avicaching points they have contributed. In both the synthetic and the real-world experiments, one action that one Avicaching participant can take is to visit a set of locations L. In practice, we assume bird watchers only choose between locations historically documented in the eBird dataset [47] hence we have information for all the locations. We assume the probability that one Avicaching participant visit a set of locations L is given by:

$$P(L) \propto \exp(w_r\theta_L^T r + w_f FL + L^T WL). \tag{23}$$

Here, we use $P(L)$ instead of $P(i)$ because each action is characterized by a set of locations. We assume L is represented as a vector of indicator variables of visited locations. $w_f FL + L^T WL$ is represented using symbol ϕ during theoretic derivation. Each column of F includes features associated with each location, such as its landscape composition, proximity to water, etc, which affects bird watchers' intrinsic utilities in visiting these locations. W is a matrix characterizing the changing of utilities when visiting multiple locations (e.g., bird watchers typically do not prefer visiting multiple locations of the same type). θ_L is a vector of indicator variables of whether visiting location set L receives each reward. w_r and w_f are the relative importance of rewards and location features. For distribution matching game, we assume $Q(L)$ has the same form as $P(L)$ although with different parameters to promote visits to remote and under-sampled locations.

Table 1: Comparison between XOR-Game$_0$ and BRQR

Size	Loss$_{XOR}$	Loss$_{BRQR}$	Time$_{XOR}$	Time$_{BRQR}$
2^2	0.0425	0.0071	147.59s	4.10s
2^4	0.0677	0.0149	154.96s	28.11s
2^8	0.0346	0.0139	196.48s	46.16s
2^{16}	0.0338	0.0178	302.24s	499.57s
2^{32}	0.0814	N/A	699.36s	>3h
2^{64}	0.0799	N/A	8476.04s	>3h

Validation on Small Games. We first validate that XOR-Game finds close-to-optimal leader's strategies on small sized games. Specifically, we focus on the zero-sum quantal response game. BRQR is used as the baseline for comparison. The difference in leader's utility values between the equilibrium and the ones found by the algorithms are shown as Loss$_{XOR}$ and Loss$_{BRQR}$ in Table 1. Here "Size" represents the number of different location sets a follower can visit, Time$_{XOR}$ and Time$_{BRQR}$ are the running times of different algorithms respectively. We can see from the table that both XOR-Game and BRQR find close-to-optimal leader's strategies. Initially XOR-Game takes longer to converge, but BRQR cannot scale to modest sized games as it runs out of a 3-hour time limit for a game with action space of $\geq 2^{32}$. XOR-Game still produces near optimal solutions in this size. Further details of this experiment (in particular, the speed the two algorithms converge to these solutions) are left to the supplementary materials.

Evaluation on Large Synthetic Benchmarks. We further evaluate the performance of XOR-Game on both the zero-sum game and the distribution matching game on large synthetic benchmarks. In these experiments, we intentionally increase the dimensionality of the reward vector r to be quadratic in the number of locations to increase the difficulty of benchmarks. To be specific, we let $\theta_L = \text{vector}(LL^T) = (l_1l_1, l_1l_2, \ldots, l_1l_n, \ldots, l_nl_n)^T$ where L is the location set vector $L = (l_1, \ldots, l_n)^T$. This is to assume one participant can receive a unique reward r_{ij} by visiting location pair (i, j). We enforce each r_{ij} to be non-negative and no greater than 1. During SGD, when r_{ij} becomes negative (or bigger than 1), we reset it to be 0 (or 1). Additional details are in the supplementary materials.

Figure 1 (left and middle) shows the performance of various algorithms as the optimization progresses. Here, each curve shows the leader's objective averaged over 20 benchmarks. For each benchmark, we let all algorithms start from the same initial solution. When computing the average, we normalize the objective function

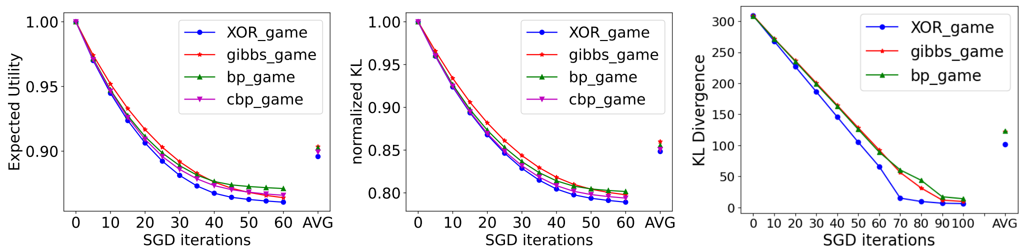

Figure 1: (Left and Middle) XOR-Game converges faster and to better solutions compared with competing approaches on synthetic datasets. (Left) Experiment on the zero sum game. (Middle) Experiment on the distribution matching game. (Right) XOR-Game converges faster than competing approaches on a behavior model learned from data collected from a real-world Avicaching game. X-axis shows the number of SGD iterations. Y-axis shows the leader's objective function L (L_0 or L_{DM}). AVG depicts $L(\overline{r_T})$.

values against that of the initial solution so each algorithm always starts from an objective function value of 1.

Notably, XOR-Game descends to the best solutions within the least number of SGD iterations for both the zero-sum games and the distribution matching games. We would like to point out that XOR-sampling in this case is efficient in obtaining the samples, even though XOR-sampling has to answer NP-complete queries. In particular, it roughly takes 1 second for XOR-sampling to obtain 100 samples during SGD, but in general it takes 50 seconds for Gibbs sampling (300 MCMC steps), 2.8 seconds for belief propagation, and 4700 seconds for chained belief propagation (cbp). Because cbp is so slow, we use 100 samples in the gradient estimation for all other approaches but only 10 samples for cbp.

Evaluation on the Avicaching Game. We then evaluate all approaches on a behavior model learned from real-world data collected from the Avicaching Game. The data comes from an actual field deployment of the Avicaching game in the eBird crowdsourcing platform between March 27 and October 29, 2015 (30 weeks) in Tompkins and Cortland counties of the New York State. A set of 50 Avicaching locations were selected, which were all publicly accessible but received no visits prior to the game. The goal of the Avicaching game was to shift the bird watchers' efforts from traditional bird watching hot spots to these Avicaching locations, harnessing Avicaching points. The numbers of Avicaching points offered for each visit to these Avicaching locations were updated every Monday. The Avicaching game was remarkably effective during this field deployment. It was reported in [54] that 19% of the bird watching effort in these two counties were shifted to these under-sampled Avicaching locations.

Before evaluating algorithms, we first learn a behavior model in the form of Equation 23 for all eBird participants. Because the field study gives an independent reward to each Avicaching location, we set θ_L to be L. $\theta_L^T r = L^T r$ represents the total reward from visiting the location set L. F includes landscape features obtained from the 2011 National Land Cover Database (NLCD). $L^T WL$ represents the change in utility functions for visiting multiple locations. Overall, $w_f FL + L^T WL$ represents the intrinsic utility of visiting locations L. Each data point consists of the set of locations L one bird watcher visits and the corresponding reward r of the week. Parameters w_r, w_f, and W are learned using Contrastive Divergence [7].

We run various algorithms for the distribution matching game to minimize the KL-divergence between the learned probability density $P(L)$ with a manually designed $Q(L)$, which promotes the visiting of under-sampled Avicaching locations and suppresses the visiting to others. The rewards were set to be greater than 0 but less than 100 for each location (same order of magnitude as the actual field deployment). Additional details in terms of learning $P(L)$ and $Q(L)$ can be found in the supplementary materials.

Figure 1(right) demonstrates that XOR-Game descends to an optimal reward allocation faster than competing approaches. All benchmarks start with identically initialized rewards. We manually inspected the solutions. The final solutions of all approaches reach almost zero for the KL-divergence, suggesting a possibility to match the learned probability density $P(L)$ to $Q(L)$ using available rewards. Nevertheless, we cannot conclude that the Avicaching game participants will act according to Q if we had the opportunities to deploy the rewards into the field. This is because all calculations are based on a learned behavior model from historical data. We cannot guarantee how much the learned model captures the subtle aspects of human decision-making and new visiting patterns may emerge as the human behavior changes with the introduction of the Avicaching game. On average, the wall-clock time for each method is: XOR_game(24h), bp_game(21h), and gibbs_game(10h). The cbp_game is exlcuded for comparison because it takes >8h per SGD iteration. In summary, XOR-Game requires the least number of SGD iterations to descend to the best leader's strategies among all benchmark algorithms. XOR-game completes in a reasonable time, has good empirical performance and provable guarantees.

5 CONCLUSION

We proposed XOR-Game to solve the convex quantal response leader-follower games with exponentially large action spaces. XOR-Game has a linear convergence speed towards the equilibrium of the leader-follower games. Our approach is based on an integration of XOR-Sampling and stochastic gradient descent, transforming the otherwise #P-hard problem into queries within the NP complexity class, while obtaining guarantees for the convergence speed. The experiments on both synthetic and real-world Avicaching games show that XOR-Game outperforms other baseline methods and hence prove its great potential for real-world applications.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their comments and suggestions. *eBird* relies on the time, dedication, and support from countless individuals and organizations. We thank the many thousands of eBird participants for their contributions and the eBird team for their support. This work was funded in part by the Leon Levy Foundation, the Wolf Creek Foundation, and the National Science Foundation (ABI sustaining: DBI-1939187, IIS-1850243, CCF-1918327).

REFERENCES

[1] Dimitris Achlioptas, Zayd Hammoudeh, and Panos Theodoropoulos. 2018. Fast and Flexible Probabilistic Model Counting. In *International Conference on Theory and Applications of Satisfiability Testing*.

[2] Dimitris Achlioptas and Panos Theodoropoulos. 2017. Probabilistic model counting with short XORs. In *International Conference on Theory and Applications of Satisfiability Testing*. Springer, 3–19.

[3] Victor Aguirregabiria and Pedro Mira. 2010. Dynamic discrete choice structural models: A survey. *Journal of Econometrics* 156, 1 (2010), 38–67. https://doi.org/10.1016/j.jeconom.2009.09.007

[4] David Belanger and Andrew McCallum. 2016. Structured prediction energy networks. In *International Conference on Machine Learning*. 983–992.

[5] Moshe Ben-Akiva and Steven R. Lerman. 1985. *Discrete choice analysis: theory and application to travel demand.* MIT Press.

[6] Michael Buro. 2003. Real-time strategy games: A new AI research challenge. In *IJCAI*, Vol. 2003. 1534–1535.

[7] Miguel Á. Carreira-Perpiñán and Geoffrey E. Hinton. 2005. On Contrastive Divergence Learning. In *AISTATS*.

[8] Jakub Černỳ, Viliam Lisỳ, Branislav Bošanskỳ, and Bo An. 2021. Computing Quantal Stackelberg Equilibrium in Extensive-Form Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5260–5268.

[9] Supratik Chakraborty, Daniel J. Fremont, Kuldeep S. Meel, Sanjit A. Seshia, and Moshe Y. Vardi. 2014. Distribution-aware Sampling and Weighted Model Counting for SAT. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*.

[10] Mark Chavira, Adnan Darwiche, and Manfred Jaeger. 2006. Compiling relational Bayesian networks for exact inference. *Int. J. Approx. Reasoning* (2006).

[11] Kam-Fung Cheung and Michael GH Bell. 2021. Attacker–defender model against quantal response adversaries for cyber security in logistics management: An introductory study. *European Journal of Operational Research* 291, 2 (2021), 471–481.

[12] Renaud Chicoisne and Fernando Ordóñez. 2016. Risk averse Stackelberg security games with quantal response. In *International Conference on Decision and Game Theory for Security*. Springer, 83–100.

[13] Vincent Conitzer and Tuomas Sandholm. 2006. Computing the optimal strategy to commit to. In *EC '06*.

[14] Fan Ding, Jianzhu Ma, Jinbo Xu, and Yexiang Xue. 2021. XOR-CD: Linearly Convergent Constrained Structure Generation. In *Proceedings of the Thirty-eighth International Conference on Machine Learning (ICML)*.

[15] Fan Ding and Yexiang Xue. 2021. XOR-SGD: Provable Convex Stochastic Optimization for Decision-making under Uncertainty. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*.

[16] Stefano Ermon, Carla P. Gomes, Ashish Sabharwal, and Bart Selman. 2013. Embed and Project: Discrete Sampling with Universal Hashing. In *Advances in Neural Information Processing Systems (NIPS)*.

[17] Stefano Ermon, Yexiang Xue, Russell Toth, Bistra N. Dilkina, Richard Bernstein, Theodoros Damoulas, Patrick Clark, Steve DeGloria, Andrew Mude, Christopher Barrett, and Carla P. Gomes. 2015. Learning Large-Scale Dynamic Discrete Choice Models of Spatio-Temporal Preferences with Application to Migratory Pastoralism in East Africa. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*.

[18] Ding Fan and Yexiang Xue. 2020. Contrastive Divergence Learning with Chained Belief Propagation. In *International Conference on Probabilistic Graphical Models*.

[19] Fei Fang, Peter Stone, and Milind Tambe. 2015. When Security Games Go Green: Designing Defender Strategies to Prevent Poaching and Illegal Fishing. In *IJCAI*.

[20] Noah Golowich, Harikrishna Narasimhan, and David C Parkes. 2018. Deep Learning for Multi-Facility Location Mechanism Design.. In *IJCAI*. 261–267.

[21] Carla P. Gomes, Ashish Sabharwal, and Bart Selman. 2006. Model Counting: A New Strategy for Obtaining Good Bounds. In *Proceedings of the 21st National Conference on Artificial Intelligence*.

[22] Carla P Gomes, Ashish Sabharwal, and Bart Selman. 2007. Near-Uniform Sampling of Combinatorial Spaces Using XOR Constraints. In *Advances in Neural Information Processing Systems*.

[23] Vassilis Argyrou Hajivassiliou. 1991. Simulation Estimation Methods for Limited Dependent Variable Models. *Handbook of Statistics* 11 (1991), 519–543.

[24] T. Hazan and A. Shashua. 2010. Norm-Product Belief Propagation: Primal-Dual Message-Passing for Approximate Inference. *Information Theory, IEEE Transactions on* 56, 12 (Dec 2010), 6294–6316. https://doi.org/10.1109/TIT.2010.2079014

[25] Joel L. Horowitz. 1993. SEMIPARAMETRIC ESTIMATION OF A WORK-TRIP MODE-CHOICE MODEL. *Journal of Econometrics* 58 (1993), 49–70.

[26] Bai Jiang, Tung-Yu Wu, Yifan Jin, Wing H Wong, et al. 2018. Convergence of contrastive divergence algorithm in exponential family. *The Annals of Statistics* 46, 6A (2018), 3067–3098.

[27] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *arXiv preprint arXiv:2106.06103* (2021).

[28] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[29] Xiaobo Li. 2018. *Convex Optimization and Online Learning: Their Applications in Discrete Choice Modeling and Pricing.* ProQuest Dissertations Publishing.

[30] Qiang Liu and Alexander Ihler. 2012. Belief Propagation for Structured Decision Making. In *Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, Corvallis, Oregon, 523–532.

[31] Xiaobo Ma, Bo An, Mengchen Zhao, Xiapu Luo, Lei Xue, Zhenhua Li, Tony TN Miu, and Xiaohong Guan. 2019. Randomized security patrolling for link flooding attack detection. *IEEE Transactions on Dependable and Secure Computing* 17, 4 (2019), 795–812.

[32] Daniel L McFadden. 1976. Quantal choice analaysis: A survey. *Annals of Economic and Social Measurement, Volume 5, number 4* (1976), 363–390.

[33] Richard D McKelvey, Andrew M McLennan, and Theodore L Turocy. 2006. Gambit: Software tools for game theory. (2006).

[34] Sanjog Misra and Sanjib K Mohanty. 2008. Estimating bargaining games in distribution channels. In *Working Paper*. Citeseer.

[35] Kevin P Murphy, Yair Weiss, and Michael I Jordan. 1999. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 467–475.

[36] Theo Offerman, Arthur Schram, and Joep Sonnemans. 1998. Quantal response models in step-level public good games. *European Journal of Political Economy* 14, 1 (1998), 89–100.

[37] Özlem Ömer. 2018. Dynamics of the us housing market: A quantal response statistical equilibrium approach. *Entropy* 20, 11 (2018), 831.

[38] Balázs Pejó and Gergely Biczók. 2020. Corona Games: Masks, Social Distancing and Mechanism Design. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19* (Seattle, WA, USA) *(COVID-19)*. Association for Computing Machinery, New York, NY, USA, 24–31. https://doi.org/10.1145/3423459.3430757

[39] A. Perrault, Bryan Wilder, Eric Ewing, Aditya Mate, Bistra N. Dilkina, and Milind Tambe. 2020. End-to-End Game-Focused Learning of Adversary Behavior in Security Games. In *AAAI*.

[40] Sainath Sanga, Venkata Sriram Siddhardh Nadendla, and Sajal K Das. 2021. Maximizing Social Welfare in Selfish Multi-Modal Routing using Strategic Information Design for Quantal Response Travelers. *arXiv preprint arXiv:2111.15069* (2021).

[41] Avi Segal, Kobi Gal, Ece Kamar, and Eric Horvitz. 2018. Optimizing Interventions via Offline Policy Evaluation: Studies in Citizen Science. In *AAAI-18: Thirty-Second AAAI Conference on Artificial Intelligence*.

[42] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. 2020. Controlvae: Controllable variational autoencoder. In *International Conference on Machine Learning*. PMLR, 8655–8664.

[43] Herbert A. Simon. 1990. *Bounded Rationality.* Palgrave Macmillan UK, London, 15–18. https://doi.org/10.1007/978-1-349-20568-4_5

[44] Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. 2018. Stackelberg Security Games: Looking Beyond a Decade of Success. In *IJCAI*.

[45] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. 2012. Random Utility Theory for Social Choice. In *NIPS*.

[46] Karunakaran Sudhir. 2001. Structural analysis of manufacturer pricing in the presence of a strategic retailer. *Marketing Science* 20, 3 (2001), 244–264.

[47] Brian L. Sullivan, Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theo Damoulas, André A. Dhondt, Tom Dietterich, Andrew Farnsworth, Daniel Fink, John W. Fitzpatrick, Thomas Fredericks, Jeff Gerbracht, Carla Gomes, Wesley M. Hochachka, Marshall J. Iliff, Carl Lagoze, Frank A. La Sorte, Matthew Merrifield, Will Morris, Tina B. Phillips, Mark Reynolds, Amanda D. Rodewald, Kenneth V. Rosenberg, Nancy M. Trautmann, Andrea Wiggins, David W. Winkler, Weng-Keen Wong, Christopher L. Wood, Jun Yu, and Steve Kelling. 2014. The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation* 169, 0 (2014), 31 – 40.

[48] Czander Tan. 2022. Double-Oracle Deep Reinforcement Learning for Handling Exponential Action Space in Sequential Stackelberg Security Games. (2022).

[49] Yasuhito Tanaka. 2018. Stackelberg type dynamic zero-sum game with leader and follower.

[50] K. Train, D. McFadden, and M. Ben-Akiva. 1987. The Demand for Local Telephone Service: A Fully Discrete Model of Residential Call Patterns and Service Choice. *RAND Journal of Economics* 18 (1987).

[51] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[52] Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky. 2003. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

[53] Haifeng Xu, Zinovi Rabinovich, Shaddin Dughmi, and Milind Tambe. 2015. Exploring Information Asymmetry in Two-Stage Security Games. In *AAAI*.

[54] Yexiang Xue, Ian Davies, Daniel Fink, Christopher Wood, and Carla P. Gomes. 2016. Avicaching: A Two Stage Game for Bias Reduction in Citizen Science. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[55] Yexiang Xue, Ian Davies, Daniel Fink, Christopher Wood, and Carla P. Gomes. 2016. Behavior Identification in Two-stage Games for Incentivizing Citizen Science Exploration. In *Proceedings of the 22nd International Conference on Principles and Practice of Constraint Programming (CP)*.

[56] Rong Yang, Christopher Kiekintveld, Fernando Ordónez, Milind Tambe, and Richard John. 2013. Improving resource allocation strategies against human adversaries in security games: An extended study. *Artificial Intelligence* 195 (2013), 440–469.

[57] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. 2003. Exploring Artificial Intelligence in the New Millennium. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter Understanding Belief Propagation and Its Generalizations, 239–269. http://dl.acm.org/citation.cfm?id=779343.779352

[58] Zhengyu Yin, Dmytro Korzhyk, Christopher Kiekintveld, Vincent Conitzer, and Milind Tambe. 2010. Stackelberg vs. Nash in security games: interchangeability, equivalence, and uniqueness. In *AAMAS 2010*.