Learning to Play Sequential Games versus Unknown Opponents*

Pier Giuseppe Sessa sessap@ethz.ch
Ilija Bogunovic ilijab@ethz.ch
Maryam Kamgarpour maryamk@ethz.ch
Andreas Krause krausea@ethz.ch
ETH Zürich

Abstract

We consider a repeated sequential game between a learner, who plays first, and an opponent who responds to the chosen action. We seek to design strategies for the learner to successfully interact with the opponent. While most previous approaches consider known opponent models, we focus on the setting in which the opponent's model is *unknown*. To this end, we use *kernel-based* regularity assumptions to capture and exploit the structure in the opponent's response. We propose a novel algorithm for the learner when playing against an adversarial sequence of opponents. The algorithm combines ideas from bilevel optimization and online learning to effectively balance between *exploration* (learning about the opponent's model) and *exploitation* (selecting highly rewarding actions for the learner). Our results include algorithm's regret guarantees that depend on the regularity of the opponent's response and scale *sublinearly* with the number of game rounds. Moreover, we specialize our approach to repeated *Stackelberg games*, and empirically demonstrate its effectiveness in a traffic routing and wildlife conservation task.

Keywords: Sequential Games, Online Learning, Gaussian Processes

1. Introduction

Several important real-world problems involve sequential interactions between two parties. These problems can often be modeled as two-player games, where the first player chooses a strategy and the second player responds to it. For example, in traffic networks, traffic operators plan routes for a subset of network vehicles (e.g., public transport), while the remaining vehicles (e.g., private cars) can choose their routes in response to that. The goal of the first player in these games is to find the optimal strategy (e.g., traffic operators seek the routing strategy that minimizes the overall network's congestion, *cf.*, Korilis et al. (1997)). Several algorithms have been previously proposed, successfully deployed, and used in domains such as urban roads (Jain et al., 2011b), airport security (Pita et al., 2009), wildlife protection (Yang et al., 2014), and markets (He et al., 2007), to name a few.

In many applications, complete knowledge of the game is not available, and thus, finding a good strategy for the first player becomes more challenging. The response function of the second player, that is, how the second player responds to strategies of the first player, is typically unknown and can only be inferred by repeatedly playing and observing the responses and game outcomes (Letchford et al., 2009; Blum et al., 2014). Consequently, we refer to the first and second players as *learner* and *opponent*, respectively. An additional challenge for the learner in such repeated games lies in facing a potentially different *type* of opponent at

*. This work has been submitted to the Neural Information Processing Systems (NeurIPS) Conference, 2020

every game round. In various domains (e.g., in security applications), the learner can even face an adversarially chosen sequence of opponent/attacker types (Balcan et al., 2015).

Motivated by these important considerations, we study a repeated sequential game against an *unknown* opponent with multiple types. We propose a novel algorithm for the learner when facing an adversarially chosen sequence of types. *No-regret* guarantees of our algorithm in these settings ensure that the learner's performance converges to the optimal one in hindsight (i.e., the idealized scenario in which the types' sequence and opponent's response function are known ahead of time). To that end, our algorithm learns the opponent's response function online, and gradually improves the learner's strategy throughout the game.

2. Problem Setup

We consider a sequential two-player repeated game between the learner and its opponent. The set of actions that are available to the learner and opponent in every round of the game are denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. The learner seeks to maximize its reward function $r(x, y)$ that depends on actions played by both players, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In every round of the game, the learner can face an opponent of different type $\theta_t \in \Theta$ that is unknown to the learner at the decision time. As the sequence of opponent's types can be chosen adversarially, we focus on randomized strategies for the learner as explained below. We summarize the protocol of the repeated sequential game as follows.
In every game round t:

1. The learner computes a randomized strategy $\mathbf{p}_t$, i.e., a probability distribution over $\mathcal{X}$, and samples action $x_t \sim \mathbf{p}_t$.
2. The opponent observes x_t and responds by selecting $y_t = b(x_t, \theta_t)$, where $b : \mathcal{X} \times \Theta \to \mathcal{Y}$ represents the opponent's *response function*.
3. The learner observes the opponent's type θ_t and response y_t, and receives reward $r(x_t, y_t)$.

The opponent's types $\{\theta_i\}_{i=1}^T$ can be chosen by an *adaptive* adversary, i.e., at round t, the type θ_t can depend on the sequence of randomized strategies $\{\mathbf{p}_i\}_{i=1}^t$ of the learner and on the previous realized actions $x_1, \dots, x_{t-1}$ (but not on the current action x_t). The goal of the learner is to maximize the cumulative reward $\sum_{t=1}^T r(x_t, y_t)$ over T rounds of the game. We assume that the learner knows its reward function $r(\cdot, \cdot)$, while the opponent's response function $b(\cdot, \cdot)$ is unknown. To achieve this goal, the learner has to repeatedly play the game and learn about the opponent's response function from the received feedback. After T game rounds, the performance of the learner is measured via the cumulative regret:

$$R(T) = \max_{x \in \mathcal{X}} \sum_{t=1}^{T} r(x, b(x, \theta_t)) - \sum_{t=1}^{T} r(x_t, y_t). \tag{1}$$

The regret represents the difference between the cumulative reward of a single best action from $\mathcal{X}$ and the sum of the obtained rewards. An algorithm is *no-regret* if $R(T)/T \to 0$ as $T \to \infty$.

Regularity assumptions. Attaining sub-linear regret is not possible in general for arbitrary response functions and domains, and hence, this requires further regularity assumptions. We consider a finite set of actions $\mathcal{X} \subset \mathbb{R}^d$ available to the learner, and a finite set of opponent's types $\Theta \subset \mathbb{R}^p$. We assume the unknown response function $b(x, \theta)$ is a member of a reproducing kernel Hilbert space $\mathcal{H}_k$ (RKHS), induced by some *known* positive-definite kernel function $k(x, \theta, x', \theta')$. RKHS $\mathcal{H}_k$ is a Hilbert space of (typically non-linear) well-behaved functions $b(\cdot,\cdot)$ with inner product $\langle \cdot,\cdot \rangle_k$ and norm $\|\cdot\|_k = \langle \cdot,\cdot \rangle_k^{1/2}$, such that $b(x,\theta) = \langle b, k(\cdot,\cdot,x,\theta)\rangle_k$ for every $x \in \mathcal{X}, \theta \in \Theta$ and $b \in \mathcal{H}_k$. The RKHS norm measures smoothness of b with respect to the kernel function k (it holds $\|b\|_k < \infty$ iff $b \in \mathcal{H}_k$). We assume a known bound $B > 0$ on the RKHS norm of the unknown response function, i.e., $\|b\|_k \leq B$. This assumption encodes the fact that similar opponent types and strategies of the learner lead to similar responses, where the similarity is measured by the known kernel function k. Most popularly used kernel functions that we also consider are linear, squared-exponential (RBF) and Matérn kernels (Rasmussen, 2003).

Our second regularity assumption is regarding the learner's reward function $r : \mathcal{X} \times \mathcal{Y} \to [0,1]$, which we assume is L_r-Lipschitz continuous with respect to $\|\cdot\|_1$.

3. Proposed Approach

The observed opponent's response can often contain some observational noise, e.g., in wildlife protection (see Section 4), we only get to observe an imprecise/inexact poaching location. Hence, instead of directly observing $b(x_t,\theta_t)$ at every round t, the learner receives a noisy response $y_t = b(x_t,\theta_t) + \epsilon_t$. For the sake of clarity, we consider the case of scalar responses, i.e., $y_t \in \mathbb{R}$, but in our full paper we also consider the case of vector-valued responses. We let $\mathcal{H}_t = \{\{(x_i,\theta_i,y_i,)\}_{i=1}^{t-1}, (x_t,\theta_t)\}$, and assume $\mathbb{E}[\epsilon_t|\mathcal{H}_t] = 0$ and ϵ_t is conditionally σ-sub-Gaussian, i.e., $\mathbb{E}\big[\exp(\zeta\epsilon_t)|\mathcal{H}_t\big] \leq \exp(\zeta^2\sigma^2/2)$ for any $\zeta \in \mathbb{R}$.

At every round t, by using the previously collected data $\{(x_i,\theta_i,y_i)\}_{i=1}^{t-1}$, we can compute a mean estimate of the opponent's response function via standard kernel ridge regression. This can be obtained in closed-form as:

$$\mu_t(x,\theta) = k_t(x,\theta)^T\big(K_t + \lambda I_t\big)^{-1}\boldsymbol{y}_t\,, \tag{2}$$

where $\boldsymbol{y}_t = [y_1,\dots,y_t]^T$ is the vector of observations, $\lambda > 0$ is a regularization parameter, $k_t(x,\theta) = [k(x,\theta,x_1,\theta_1),\dots,k(x,\theta,x_t,\theta_t)]^T$ and $[K_t]_{i,j} = k(x_i,\theta_i,x_j,\theta_j)$ is the kernel matrix. The variance of the proposed estimator can be obtained as:

$$\sigma_t^2(x,\theta) = k(x,\theta,x,\theta) - k_t(x,\theta)^T\big(K_t + \lambda I_t\big)^{-1}k_t(x,\theta)\,. \tag{3}$$

Moreover, we can use (2) and (3) to construct upper and lower confidence bound functions:

$$\mathrm{ucb}_t(x,\theta) := \mu_t(x,\theta) + \beta_t\sigma_t(x,\theta), \quad \mathrm{lcb}_t(x,\theta) := \mu_t(x,\theta) - \beta_t\sigma_t(x,\theta)\,, \tag{4}$$

respectively, for every $x \in \mathcal{X}, \theta \in \Theta$, where β_t is a confidence parameter. A standard result from Abbasi-Yadkori (2013); Srinivas et al. (2010) shows that under our regularity assumptions, β_t can be set such that, with high probability, response $b(x,\theta) \in [\mathrm{lcb}_t(x,\theta), \mathrm{ucb}_t(x,\theta)]$ for every $(x,\theta) \in \mathcal{X} \times \Theta$ and $t \geq 1$.

Before moving to our main results, we define a sample complexity parameter that quantifies the *maximum information gain* about the unknown function from noisy observations:

$$\gamma_t := \max_{\{(x_i,\theta_i)\}_{i=1}^t} 0.5 \log\det(I_t + K_t/\lambda). \tag{5}$$

It has been introduced by Srinivas et al. (2010) and later on used in various theoretical works on Bayesian optimization. Analytical bounds that are sublinear in t are known for popularly used kernels (Srinivas et al., 2010), e.g., when $\mathcal{X} \times \Theta \subset \mathbb{R}^d$, we have $\gamma_t \leq \mathcal{O}(\log(t)^{d+1})$ and $\gamma_t \leq \mathcal{O}(d\log(t))$ for squared exponential and linear kernels, respectively. This quantity characterizes the regret bounds obtained in the next sections.

noend 1 The StackelUCB algorithm

Input: Finite action set $\mathcal{X} \subset \mathbb{R}^d$, kernel $k(\cdot,\cdot)$, param. $\lambda, \{\beta_t\}_{t\geq 1}, \eta$

1: Initialize: Uniform strategy $\mathbf{p}_1 = \frac{1}{|\mathcal{X}|}\mathbf{1}_{|\mathcal{X}|}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Sample action $x_t \sim \mathbf{p}_t$ // Opponent θ_t observes x_t and computes $b(x_t, \theta_t)$
4: Observe θ_t and noisy response $y_t = b(x_t, \theta_t) + \epsilon_t$
5: Compute optimistic reward estimates:
$\forall x \in \mathcal{X} :\ \tilde{r}_t(x, \theta_t) := \max_y r(x, y), \quad \text{s.t.} \quad y \in \big[\mathrm{lcb}_t(x, \theta_t), \mathrm{ucb}_t(x, \theta_t)\big]$
6: Perform strategy update: $\forall x \in \mathcal{X} :\ \mathbf{p}_{t+1}[x] \propto \mathbf{p}_t[x] \cdot \exp\big(\eta \cdot \tilde{r}_t(x, \theta_t)\big)$
7: Update: μ_{t+1}, σ_{t+1} with $\{(x_t, \theta_t, y_t)\}$ (via (2), (3)), and $\mathrm{ucb}_{t+1}, \mathrm{lcb}_{t+1}$ (via (4))

3.1 The StackelUCB Algorithm

The considered problem (Section 2) can be seen as an instance of adversarial online learning (Cesa-Bianchi and Lugosi, 2006) in which an adversary chooses a reward function $r_t(\cdot)$ in every round t, while the learner (without knowing the reward function) selects action x_t and subsequently receives reward $r_t(x_t)$. To achieve *no-regret*, the learner needs to maintain a probability distribution $\mathbf{p}_t$ over the set $\mathcal{X}$ of available actions and play randomly according to it. *Multiplicative Weights* (MW) (Littlestone and Warmuth, 1994) algorithms such as Exp3 (Auer et al., 2003) and Hedge (Freund and Schapire, 1997) are popular no-regret methods for updating $\mathbf{p}_t$, depending on the feedback available to the learner in every round. The former only needs observing reward of the played action $r_t(x_t)$ (*bandit* feedback), while the latter requires access to the entire reward function $r_t(\cdot)$ at every t (*full-information* feedback).

The considered game setup corresponds (from the learner's perspective) to the particular online learning problem in which $r_t(\cdot) := r(\cdot, b(\cdot, \theta_t))$, type θ_t is revealed, and the bandit observation y_t is observed by the learner. Full-information feedback, however, is not available as $b(\cdot, \theta_t)$ is unknown. To alleviate this, similarly to Sessa et al. (2019), we compute "optimistic" reward estimates to emulate the full-information feedback. Based on previously observed data, we establish upper and lower confidence bounds $\mathrm{ucb}_t(\cdot)$ and $\mathrm{lcb}_t(\cdot)$, of the opponent's response function via (4). These are then used to estimate the optimistic rewards of the learner for any $x \in \mathcal{X}$ at round t as:

$$\tilde{r}_t(x, \theta_t) := \max_y \ r(x, y) \quad \text{s.t.} \quad y \in \big[\mathrm{lcb}_t(x, \theta_t), \mathrm{ucb}_t(x, \theta_t)\big]. \tag{6}$$

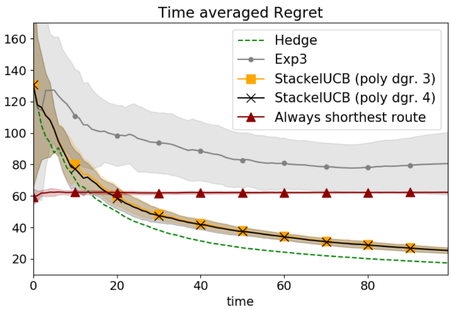
Optimistic rewards allow the learner to control the maximum incurred regret, while Lipschitness of $r(.)$ ensures that learning the opponent's response function (via (2) and (3)) translates to more accurate reward estimates. The proposed approach is summarized in our novel StackelUCB algorithm (see Algorithm 1) which provides the following guarantee.

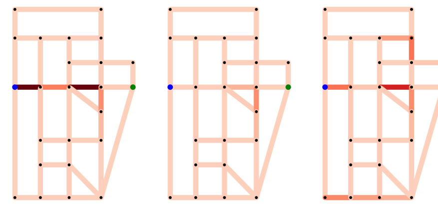
Theorem 1 *For any $\delta \in (0,1)$, the regret of* StackelUCB *when used with $\lambda \geq 1$, $\beta_t = \sigma\lambda^{-1}\sqrt{2\log\left(\frac{1}{\delta}\right) + \log(\det(I_t + K_t/\lambda))} + \lambda^{-1/2}B$, and learning step $\eta = \sqrt{8\log(|\mathcal{X}|)/T}$, is bounded, with probability at least $1 - 2\delta$, by*

$$R(T) \leq \sqrt{\tfrac{1}{2}T\log|\mathcal{X}|} + \sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)} + \ 4L_r\beta_T\sqrt{T\lambda\gamma_T},$$

where $B \geq \|b\|_{\mathcal{H}_k}$ and γ_T is the maximum information gain defined in (5).

The obtained regret bound scales sublinearly with T, and depends on the regret obtained from playing Hedge (first two terms) and learning of the opponent's response function (last





	Shortest route	0% routed	StackelUCB
Avg. congestion	15.97	1.03	3.51
Cumul. reward	21'645.4	-813.5	25'330.5

Figure 1: **Left:** Time-averaged regret of the operator using different routing strategies. StackelUCB (polynomial kernels of degree 3 or 4) leads to a smaller regret compared to the considered baselines and performs comparably to the idealized Hedge algorithm. **Right:** Edges' congestion (color intensity proportional to the time-averaged congestion) when the operator at each round: (left) Routes 100% of the units via the shortest route, (middle) Routes 0% of units, and (right) Uses StackelUCB.

term in the regret bound). We note that Exp3 attains $\mathcal{O}(\sqrt{T|\mathcal{X}|\log|\mathcal{X}|})$ while Hedge attains improved $\mathcal{O}(\sqrt{T\log|\mathcal{X}|})$ regret bound which scales favourably with the number of available actions $|\mathcal{X}|$. The same holds for our algorithm, but crucially – unlike Hedge – our algorithm uses the bandit feedback only.

3.2 Single Opponent Type

In case the learner is playing against an opponent of a *single* type $\bar{\theta}$, in our full paper we show that a simpler version of StackelUCB achieves an improved regret bound of $4L_r\beta_T\sqrt{T\lambda\gamma_T}$. The strategy consists, at each time t, of selecting $x_t = \arg\max_{x\in\mathcal{X}} \; \tilde{r}_t\big(x,\bar{\theta}\big)$ and is reminiscent of the GP-UCB algorithm used in standard Bayesian optimization (Srinivas et al., 2010).

3.3 Learning in Repeated Stackelberg Games

Repeated Stackelberg games (von Stackelberg, 1934) are sequential games between a leader (learner) and a follower (opponent). They can be mapped to our setup of Section 2 by letting $\mathcal{X} = \Delta^{n_l}$ be the leader decision set, where Δ^{n_l} stands for n_l-dimensional simplex. Moreover, the opponent's response function in a Stackelberg game assumes the specific *best-response* form $b(x_t,\theta_t) = \arg\max_{y\in\mathcal{Y}} U_{\theta_t}(x_t,y)$, where $U_{\theta_t}(x,y)$ represents the expected utility of the follower of type θ_t under the leader's strategy x. Balcan et al. (2015) shows that a regret bound of $\mathcal{O}\big(\sqrt{T\cdot \mathrm{poly}(n_l,n_f,k_f)}\big)$ (n_f and k_f are the numbers of actions available to the follower and possible follower types, respectively) can be achieved assuming a finite set of followers with *known* utilities. In this work, we show that StackelUCB leads to a regret of $\mathcal{O}\big(\sqrt{Tn_l\log(L_rL_b\sqrt{n_lT})} + L_r\beta_T\sqrt{T\lambda\gamma_T}\big)$ in the more challenging setting where such utilities are *unknown* (also, potentially with an infinite number of types).

4. Experiments

We evaluate the proposed algorithms in traffic routing and wildlife conservation tasks.

Routing Vehicles in Congested Traffic Networks. We consider a traffic routing task in the network of Sioux-Falls (LeBlanc et al., 1975), in which the goal of the network operator is to route 300 units (e.g., a fleet of autonomous vehicles) between the two nodes of the network (depicted as blue and green nodes in Figure 1). At the same time, the goal of the operator is to avoid the network becoming overly congested. We model this problem as a repeated sequential game (as defined in Section 2) between the network operator (learner)

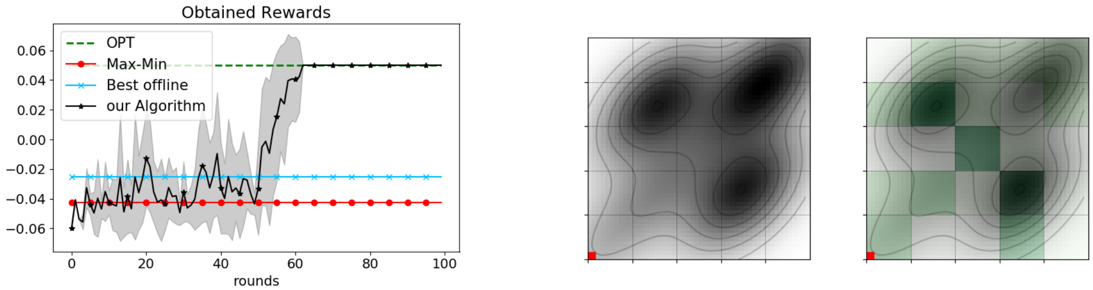


Figure 2: **Left:** Obtained rewards when the rangers know the poachers' model (OPT), or use different patrol strategies. Our algorithm discovers the optimal strategy in ∼60 rounds and outperforms the considered baselines. **Right:** Park animal density (left plot) and rangers' patrol strategy (right plot, where probabilities are proportional to the green color intensity) computed with our algorithm. The poachers' model and starting location (red square) are not known by the rangers ahead of time.

and the rest of the users present in the network (opponent), where users' preferences and the network's congestion model are unknown to the operator.

In Figure 1 we compare the performance of the network operator when using StackelUCB to select routes with 1) routing 100% of the units via the shortest route at every round, 2) routing 0% of the units at every round, 3) the Exp3 algorithm and 4) the Hedge algorithm. StackelUCB leads to a significantly smaller regret compared to the considered baselines (the regret of baseline 2 is above the y-axis limit), and its performance is comparable to the full-information Hedge algorithm. Moreover, we report the cumulative reward obtained by the operator when using StackelUCB and other two baselines, together with the resulting time-averaged congestion levels. The network's average congestion is very low when 0% of the units are routed, while the central edges become extremely congested when 100% of the units are routed via the shortest route. Instead, the proposed game model and StackelUCB algorithm allow the operator to select alternative routes depending on the users' demands, leading to improved congestion and a larger cumulative reward compared to the baselines.

Wildlife Protection against Poaching Activity. We consider a wildlife conservation task where the goal of park rangers is to protect animals from poaching activities. We model this problem as a sequential game between the rangers, who commit to a patrol strategy x (i.e., covering each park area with some probability), and the poachers that observe the rangers' strategy to decide upon a poaching location $y = b(x)$. We use the game model of Kar et al. (2015) to define poachers' model $b(\cdot)$ and rangers' reward function $r(\cdot)$. We study the repeated version of this game in which the rangers start with no information about the poachers' model and use the algorithm discussed in Section 3.2 to discover the best patrol strategy online.

In Figure 2 (left plot), we compare the performance of our algorithm with the ones achieved by: 1) Optimal strategy (OPT) $x^\star = \arg\max_{x \in \mathcal{X}} r(x, b(x))$ with known poachers' model, 2) Max-Min, i.e, $x_{\mathrm{m}} = \arg\max_{x \in \mathcal{X}} \min_y r(x, y)$, which assumes the worst possible poaching location, and 3) Best-offline, that is, $x_{\mathrm{o}} = \arg\max_{x \in \mathcal{X}} r(x, \mu_{\mathrm{o}}(x))$, where $\mu_{\mathrm{o}}(\cdot)$ is the mean estimate of $b(\cdot)$ computed *offline* as in (2) by using 1'000 random data points. Our algorithm outperforms the considered baselines and discovers the optimal patrol strategy after $\sim$ 60 rounds. We depict the discovered strategy in Figure 2 (rightmost plot). We observe that the cells covered with higher probabilities are the ones with a high animal density near to the poachers' starting location (despite the latter is unknown to the algorithm).

References

Yasin Abbasi-Yadkori. Online learning for linearly parametrized control problems. 2013.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.*, 32(1):48–77, January 2003.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Commitment Without Regrets: Online Learning in Stackelberg Security Games. In *ACM Conference on Economics and Computation (EC)*, 2015.

Lorenzo Bisi, Giuseppe De Nittis, Francesco Trovò, Marcello Restelli, and Nicola Gatti. Regret Minimization Algorithms for the Followers Behaviour Identification in Leadership Games. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Learning Optimal Commitment to Overcome Insecurity. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

Andreea Bobu, Dexter R. R. Scobee, Jaime F. Fisac, S. Shankar Sastry, and Anca D. Dragan. LESS is More: Rethinking Probabilistic Models of Human Behavior. 2020.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

Nando de Freitas, Alex Smola, and Masrour Zoghi. Regret bounds for deterministic Gaussian process bandits. *ArXiv*, abs/1203.2177, 2012.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimilano Pontil. Bilevel Programming for Hyperparameter Optimization and Meta-Learning. *ArXiv*, abs/1806.04910, 2018.

Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Víctor Gallego, Roi Naveiro, David Ríos Insua, and David Gomez-Ullate Oteiza. Opponent Aware Reinforcement Learning. *ArXiv*, abs/1908.08773, 2019.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé. Opponent Modeling in Deep Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 2016.

Xiuli He, Ashutosh Prasad, Suresh P. Sethi, and Genaro J. Gutierrez. A survey of Stackelberg differential game models in supply and marketing channels. *Journal of Systems Science and Systems Engineering*, 16(4):385–413, 2007.

Manish Jain, Christopher Kiekintveld, and Milind Tambe. Quality-bounded solutions for finite Bayesian Stackelberg games: scaling up. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011a.

Manish Jain, Dmytro Korzhyk, Ondřej Vaněk, Vincent Conitzer, Michal Pěchouček, and Milind Tambe. A Double Oracle Algorithm for Zero-Sum Security Games on Graphs. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011b.

Debarun Kar, Fei Fang, Francesco Maria Delle Fave, Nicole D. Sintov, and Milind Tambe. "A Game of Thrones": When Human Behavior Models Compete in Repeated Stackelberg Security Games. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2015.

Debarun Kar, Benjamin J. Ford, Shahrzad Gholami, Fei Fang, Andrew J. Plumptre, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Mustapha Nsubaga, and Joshua Mabonga. Cloudy with a Chance of Poaching: Adversary Behavior Modeling and Forecasting with Real-World Poaching Data. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2017.

Yannis A. Korilis, Aurel A. Lazar, and Ariel Orda. Achieving Network Optima Using Stackelberg Routing Strategies. *IEEE/ACM Trans. Netw.*, 5(1):161–173, 1997.

Larry J. LeBlanc, Edward K. Morlok, and William P. Pierskalla. An efficient approach to solving the road network equilibrium traffic assignment problem. In *Transportation Research Vol. 9*, pages 309–318, 1975.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and Approximating the Optimal Strategy to Commit To. In *International Symposium on Algorithmic Game Theory (SAGT)*, 2009.

D. Liao-McPherson, M. Huang, and I. Kolmanovsky. A Regularized and Smoothed Fischer–Burmeister Method for Quadratic Programming With Applications to Model Predictive Control. *IEEE Transactions on Automatic Control*, 64(7):2937–2944, 2019.

N. Littlestone and M.K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108(2):212 – 261, 1994.

Janusz Marecki, Gerry Tesauro, and Richard Segal. Playing Repeated Stackelberg Games with Unknown Opponents. In *International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2012.

Thanh H. Nguyen, Rong Yang, Amos Azaria, Sarit Kraus, and Milind Tambe. Analyzing the Effectiveness of Adversary Modeling in Security Games. In *AAAI Conference on Artificial Intelligence*, 2013.

Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordóñez, and Sarit Kraus. Playing games for security: an efficient exact algorithm for solving Bayesian Stackelberg games. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2008.

Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning Optimal Strategies to Commit To. In *AAAI Conference on Artificial Intelligence*, 2019.

James Pita, Manish Jain, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Using Game Theory for Los Angeles Airport Security. *AI Magazine*, 30:43–57, 2009.

Roberta Raileanu, Emily L. Denton, Arthur Szlam, and Rob Fergus. Modeling Others using Oneself in Multi-Agent Reinforcement Learning. *ArXiv*, abs/1802.09640, 2018.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. No-Regret Learning in Unknown Games with Correlated Payoffs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

Arunesh Sinha, Debarun Kar, and Milind Tambe. Learning Adversary Behavior in Security Games: A PAC Model Perspective. In *International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, 2016.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, 2010.

Zheng Tian, Ying Wen, Zhichen Gong, Faiz Punakkath, Shihao Zou, and Jun Wang. A Regularized Opponent Model with Maximum Entropy Objective. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

H. von Stackelberg. *Marktform und Gleichgewicht*. Die Handelsblatt-Bibliothek "Klassiker der Nationalökonomie". J. Springer, 1934.

Rong Yang, Benjamin J. Ford, Milind Tambe, and Andrew Lemieux. Adaptive resource allocation for wildlife protection against illegal poachers. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2014.