Adversarial Forecasting through Adversarial Risk Analysis within a DDDAS Framework

Tahir Ekin

Gregg Endowed Associate Professor of Analytics

Department of Information Systems & Analytics
McCoy College of Business, Texas State University

joint paper with Roi Naveiro and Jose Manuel Camacho Rodriguez (ICMAT-CSIC)

Dynamic Data Driven Applications Systems (DDDAS) 2022 Conference
Cambridge MA

October 2022

Adversarial Forecasting: Background

Standard assumptions of forecasting

- Clean and legitimate data streams
- Identically distributed training and test data

But what about

- Potential manipulation of digital data streams to influence forecasts
- Attack on forecasting output through manipulating input data (data-fiddler attack) or model parameters (structural attack)
- Attempts of pushing the forecast towards (i.e., attractive attack) or away from a certain region of interest (i.e., repulsive attack).

How to incorporate the impact of such manipulations and the associated uncertainty and incomplete knowledge into forecasting models?

Adversarial Forecasting: Examples

- Military when an adversary attempts to poison the input data to alter forecasts and automated decisions
- Accounting where the attacker cooks the books by misreporting past values to avoid an audit
- E-commerce recommendation systems: botnets to manipulate the number of visits to a web site which would be inputted into web traffic prediction and ad placement models
- Electricity load management, demand response models: Attacker corrupts the past price or load
- Assisted driving systems: manipulating weather data of a sensor

Adversarial Forecasting: Literature

- Adversarial machine learning: mostly adversarial classification, e,g., Dalvi et al. (2004)
- Emerging literature of temporal and unsupervised adversarial learning
    - Stackelberg games between the learner (predictive method) and the adversary assuming common knowledge about adversary's costs and action space: Bruckner and Scheffer (2011)
    - Poisoning of trained linear autoregressive forecasting models where the attacker manipulates the inputs to drive the latent space of a linear AR model towards a region of interest: Alfeld et al. (2016)
    - Load forecasting, attacker injecting malicious data in temperature from online weather forecast APIs : Chen et al. (2019)
    - Sequential adversarial attacks on Kalman Filter output which is fed into the forward collision warning system: Ma et al. (2020)

Adversarial attacks within a decision analysis framework

- Gaps in literature:
  - Limited adversarial statistical theory and computational algorithms: Adversarial forecasting and unsupervised methods
  - Common knowledge assumption in existing adversarial games: Incomplete information with uncertainty
- Goal: Utilize Bayesian decision theory (adversarial risk analysis- ARA) principles in developing the theoretical, computational and applicable frameworks for adversarial forecasting embedded to decision making
- Potential practical needs: Incomplete information inherent in adversarial settings, more than one decision maker, non-cooperative dynamic decision environments, (decision dependent) uncertainty

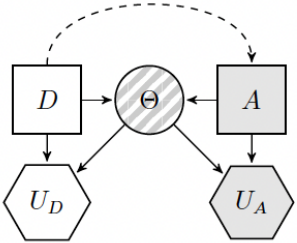
- Focus is on incorporating the impact of data-fiddler grey-box attacks and the associated uncertainty and incomplete information into (defender's) forecasting models

DDDAS: Relevant background

- DDDAS: systems that utilize physical sensor/measurement data to dynamically update a computational model (Darema, 2004)
- Dynamic feedback loop
- Applicability in decision support applications with adversarial contexts
- Attack detection via anomaly detection modules (Combita et al., DDDAS18), outlier detection based defenses against data poisoning attacks for classifiers (Li et al., DDDAS20), iterative dynamic data repair in sensor networks for power network load forecasting models (Zhou et al., DDDAS20)
- rDDDAS to operate safely in a compromised environment while building tolerating defenses (Dsouza et al., 2013) "It is almost impossible to build perfectly secure cyber systems and fully avoid the impact of adversarial attacks."
- Resilient adaptive machine learning ensemble that tolerates adversarial learning attacks via moving target defense (Yao et al., DDDAS20)

ARA (Rios Insua, Rios, Banks, JASA 2009)

- Models games as a decision-theoretic problem from the expected utility maximizing perspective of a given player.
- Relaxes common knowledge assumption and the common prior hyp.



- Defender first solving Attacker's problem while incorporating her uncertainty about his probabilities and utilities (hence optimal decision) by using $p_D(a|d) = P_F[A^*(d) = a]$; then solving her own problem as in $d^*_{\text{ARA}} = argmax_d \int \psi_D(d,a) p_D(a|d) da$
- ARA for AML: Naveiro et al. (2019), Gonzalez-Ortega et al. (2021)

Adversarial ARIMA within DDDAS for Load forecasting

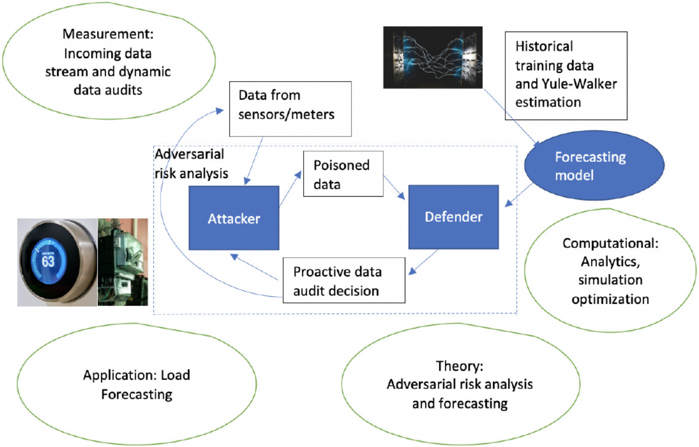
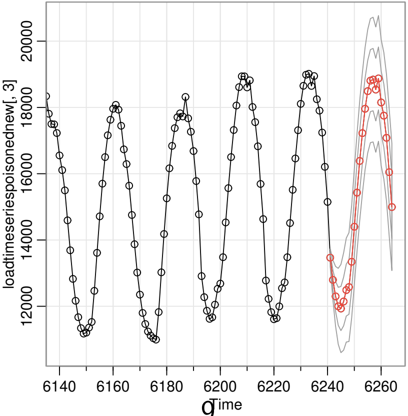


Fig. 1. Proposed DDDAS framework

Data stream for demonstration

- Data from Electric Reliability Council of Texas (ERCOT)
- Focus: Time series of 2015-2017 summer electricity hourly load of Houston
- High seasonality peaking around 16-17, and bottoming around 4-5.
- The best fit by a seasonal SARIMA $(5, 0, 0)x(2, 1, 0)^{24}$ model



Adversarial ARIMA: Attacker's decision model

- Determine the perturbation size,$\boldsymbol{\eta}$, that leads to decreased load forecasts (which could eventually motivate Defender decrease supply leading to outages)
- $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\eta}$
-

$$\max_{\eta} \qquad \psi_A(\boldsymbol{\eta}, \beta, \Phi) = \sum_{h=1}^{h=H} [u(f^h | \boldsymbol{X}) - u(f^h | \boldsymbol{X}, \boldsymbol{\eta})]$$

$$\eta_t < L \quad \forall t, \sum_{t=1}^{t=T} c_t * \eta_t <= B, \boldsymbol{\eta} \in A(\boldsymbol{\eta}, \beta)$$

- Defender's decision of data auditing, β of time t is denoted as $\beta = t$ which indicates $\eta_t = 0$ updating the Attacker's set of achievable attacks, $A(\boldsymbol{\eta}, \beta)$

Adversarial ARIMA: Attacker's decision model

- ARA allows the Attacker to acknowledge his uncertainty about Defender's forecasting model, $p_D(\Phi)$ as well as the data audit decision, β affecting $A(\boldsymbol{\eta}, \beta)$
- Attacker's expected utility:

$$\psi_A(\boldsymbol{\eta}) = \int \left[\int u_A(\boldsymbol{\eta}, \beta, \Phi) \, p_A(\Phi | \boldsymbol{\eta}, \beta) d\Phi \right] p_A(\beta | \boldsymbol{\eta}) d\beta$$

- To find $p_A(\beta|\boldsymbol{\eta})$, he would induce a distribution over the Defender's expected utility $\psi_D(\boldsymbol{\eta})$, from which random optimal alternative, $\beta^*(\boldsymbol{\eta})$ is computed for each $\boldsymbol{\eta}$ decision alternative.
- $\boldsymbol{\eta}^*_{\text{ARA}} = argmax_{\boldsymbol{\eta} \in A(\boldsymbol{\eta}, \beta)} \psi_A(\boldsymbol{\eta})$.

Adversarial ARIMA: Defender's decision model

- Defender's problem to impact $A(\boldsymbol{\eta})$ via data audits $\beta = t$ with the goal of minimizing the maximum potential attack impact
- After Defender retrieves the data (which could have been poisoned) and makes a forecast, that is compared to a predetermined threshold within a simple decision support tool.
- $$\min_{\beta} \quad max \sum_{h=1}^{h=H} [PT^h - u(f^h | \boldsymbol{Y}, \beta)]$$
- If the load forecast is less than the predetermined threshold for h^{th} time period, PT^h (mean historical load), there may be supply adjustments.

Adversarial ARIMA: Insights

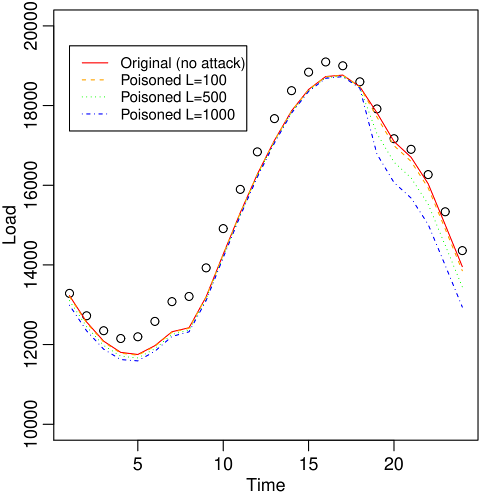


Figure: Forecasts with varying levels of attack

Computational Algorithms

Need for computational advances in solution methods

$$\max_x E_\xi[Q(x,\xi)]$$

- Sequential estimation of the expectation and optimization
- Estimation challenges with high dimensions, skewed and/or multi-modal and/or decision dependent uncertainty
- Optimization challenges with complex objective functions
- Customized solution approaches: ranking & selection heuristics, augmented probability simulation

Augmented probability simulation

$$\max_{z} \sum_{\rho} \int_{\mathcal{A}} \int_{\mathcal{B}} \int_{\Pi} u_Z(z,\phi) g_A(A) g_B(B) g_\pi(\pi) P_\rho(\rho) \pi B A$$

- Ekin et al. (2022) EJOR
- APS converts this into a grand simulation problem by simultaneously performing expectation and optimization

$$\breve{g}(z, A, B, \pi, \rho) \propto u_Z(z,\phi) g_A(A) g_B(B) g_\pi(\pi) P_\rho(\rho).$$

- When we sample $(z, A, B, \pi, \rho) \sim \breve{g}(z, A, B, \pi, \rho)$, the mode of z samples approximates the optimal solution.
- Transformation

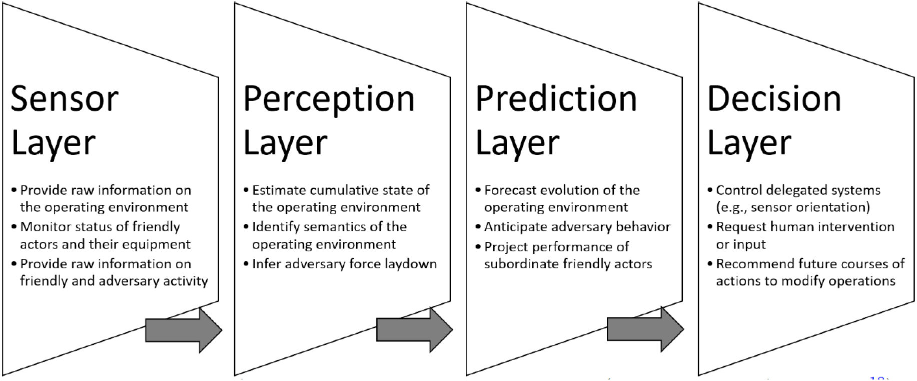
$$\breve{g}_H(z, \{A^h, B^h, \pi^h, \rho^h\}_{h\in\mathcal{H}}) \propto \prod_{h\in\mathcal{H}} u_Z\left(z, \phi^h\right) g_A(A^h) g_B(B^h) g_\pi(\pi^h) P_\rho(\rho^h)$$

- Metropolis within Gibbs

$$\breve{g}_H(z_t | z_{-t}, \{A^h, B^h, \pi^h, \rho^h\}_{h\in\mathcal{H}}) \propto \exp\left(\sum \log \left[u_Z\left(z_t \cup z_{-t}, \phi^h \right) \right] \right).$$

Application Areas

- Application use cases/examples: Electricity load management, demand response models, cybersecurity, command and control



Sensor Layer

- Provide raw information on the operating environment
- Monitor status of friendly actors and their equipment
- Provide raw information on friendly and adversary activity

Perception Layer

- Estimate cumulative state of the operating environment
- Identify semantics of the operating environment
- Infer adversary force laydown

Prediction Layer

- Forecast evolution of the operating environment
- Anticipate adversary behavior
- Project performance of subordinate friendly actors

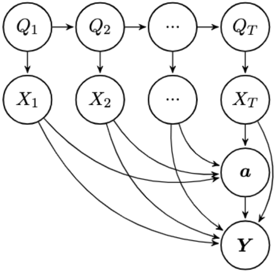
Decision Layer

- Control delegated systems (e.g., sensor orientation)
- Request human intervention or input
- Recommend future courses of actions to modify operations

Figure: Conceptual architecture of multi-domain C2 systems (Caballero, Friend and Blasch (2021))

Relevant work: HMM poisoning

- Ex: Satellite data transmission to down link ground station
- Batch data: full observation sequence $\{x_t\}_{t \in \mathcal{T}}$
- Grey-box attack: knows family (HMM), but not the parameterization
- Attacks with uncertain outcomes and (deterministic) cost coupled with risk of discovery



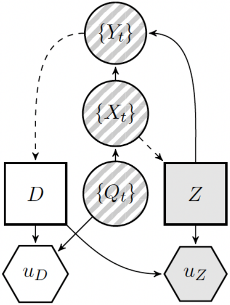


Figure 3: BAID for HMM Poisoning Problem

Concluding remarks

- Major takeaways
  - Theoretical focus on unsupervised learning and classical statistical forecasting methods
  - Incomplete information and uncertainty in an adversarial environment for more than one decision maker
  - Proactive and dynamic protection for resilient grids
- What is next?
  - Potential applications to C2 problem sets: e.g. USAF's Air Operations Centers, decision models with automation that use sensor data
  - Implications for defender and solving defender's problems
  - Implications for autonomous systems and robustifying forecasts
  - Accommodation of multi-sensor data under adversarial attacks, multiple decision makers of different types
  - Theoretical extensions to structural attacks and other methods

Disclaimer and Acknowledgment

- This material is based upon work supported by the Air Force Office of Scientific Research award FA9550-19-S-0003, Air Force Office of Scientific Research European Office of Aerospace Research and Development award FA8655-21-1-7042, and Steven R. “Steve” Gregg Excellence Endowed Professorship.

- Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense and Air Force Office of Scientific Research.

Thank you

Contact: tahirekin@txstate.edu