Online Resource Allocation with Noisy Predictions

Jessica Maghakian

Department of Applied Mathematics and Statistics, Stony Brook University

Homepage: `http://www.ams.sunysb.edu/~jmaghaki`

Thesis: Online Resource Allocation with Noisy Predictions

Advisor: Zhenhua Liu, Department of Applied Mathematics and Statistics, Stony Brook University

Brief Biography: Jessica Maghakian is a final-year PhD candidate in Operations Research at Stony Brook University. She has collaborated with several industry partners and interned at Microsoft Research NYC. Jessica's research combines data-driven decision making with traditional online algorithms approaches to solve challenging resource allocation problems that arise in IT infrastructure systems. Her work has been awarded an NSF Graduate Research Fellowship, a Stony Brook University STRIDE Fellowship for excellence in visualization, science communication and decision-support, and a Best Paper Nomination at ACM e-Energy. She received her Bachelors of Science from the Massachusetts Institute of Technology (MIT) with a double major in Mathematics and Brain and Cognitive Sciences.

Research Summary: By as early as next year, 40% of infrastructure and operations teams in large organizations will use AI-augmented solutions. The rapid adoption of AI and ML in infrastructure presents both a challenge and an opportunity for online optimization algorithm design. On the one hand, IT infrastructure requires resource allocation algorithms that perform well even in worst-case settings. Although classical online algorithms offer theoretical performance guarantees, their conservative decision making is not suitable for typical production inputs. On the other hand, black box ML predictions generated from historical data can be highly accurate, however these predictions are imperfect and their quality can also unexpectedly degrade with little warning.

During my dissertation, I designed *learning-augmented online algorithms* with a focus on providing theoretical guarantees as well as strong practical performance. The infrastructure setting provides many unique difficulties and obstacles. Some of the technical challenges I have focused on include:

- Cloud resource managers serve **non-stationary demands** that are notoriously unpredictable and volatile. Existing methods that jointly choose prediction and allocation algorithms perform poorly in this non-stationary regime. I designed an adaptive meta-algorithm outperforms state-of-the-art competitors in 71% of trials on highly non-stationary Microsoft Azure virtual machine (VM) demand traces.
- Heterogeneity across users populations can result in black-box **predictions with varying noise levels**. Data-driven approaches fail when predictions are very noisy, while classical online algorithms cannot exploit situations when predictions are accurate. I designed algorithms that provide guaranteed strong performance in the average-case and worst-case, independently of prediction quality.
- System operators have access to **different types of predictors** (e.g. predictions of aggregated monthly demand statistics, fine-grained predictions over individual VMs) yet online optimization with multiple predictors remains under-researched. I designed algorithms that achieve significant theoretical and empirical performance improvements over existing methods by leveraging different types of noisy predictors.

There is a vast body of literature on incorporating predictions for resource allocation. However, the majority of existing work exclusively considers predictions of *one quantity*. There is a significant gap between the work done by the algorithms community and real-world systems, where operators have access to large amounts of high quality historical data that can be used to generate predictions of many *different quantities*. In machine learning, sets of weak learners can be combined to create a strong learner via boosting. My research explores the benefits of incorporating multiple noisy predictors into online optimization. In particular, I examine whether it is possible to maintain worst-case theoretical performance guarantees, for arbitrary noise models. In the following sections, I introduce different application-motivated problems and algorithms that can leverage different noisy predictors to outperform existing solutions.

1. *Noisy predictions in a non-stationary environment*

Much of the online algorithms with prediction literature has focused on leveraging *input predictors*, which predict the unknown future inputs (e.g. future user demands) of the online algorithm. Since prediction and control algorithms are typically studied independently by different research communities, oftentimes the combination of best input predictor and best control algorithm might not be the best possible pairing. Finding the best pairing of prediction and control algorithm is the *online algorithm selection problem*. Although many meta-algorithms have been developed

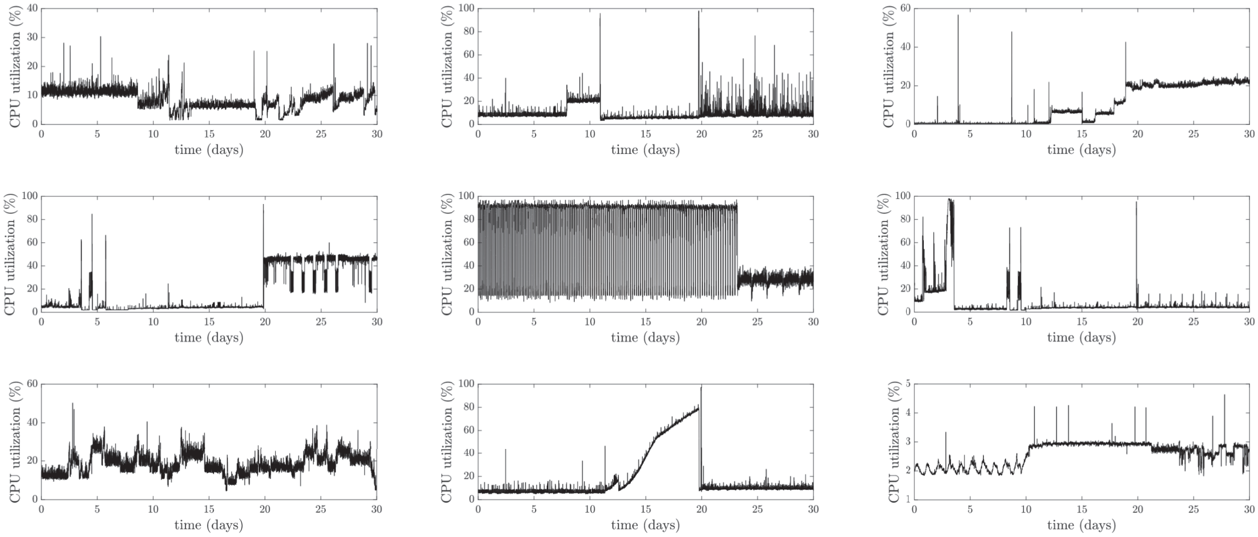

Figure 1: Average CPU utilization for nine different Microsoft Azure virtual machines over a one month contract period. The traces exhibit abrupt shifts in demand characteristics, requiring novel aperiodic algorithms.

for this problem, such as Follow the Leader and Weighted Majority, they are not suitable in cloud systems due to (1) *non-stationary* resource demands that are unpredictable and volatile over the long service contract periods and (2) *time-coupling* due to the startup and shutdown cost of servers, virtual machine migration and data transfer.

I tackled the problem of prediction and control algorithm selection in the context of resource provisioning in the cloud [5]. First, I exemplified the need to address non-stationarity by showcasing 200+ traces from the Microsoft Azure Public Dataset (see Figure 1). Rather than gradual distributional drift, the traces exhibit more aggressive and sudden changes. I developed a meta-algorithm which combines online change point detection with a modified weighted majority approach that can accommodate the switching costs arising from time-coupling. By detecting when a shift in the statistical profile of user demand occurs, and then switching to a more optimal prediction and allocation algorithm pairing, the meta-algorithm flexibly handles even the most volatile demands. My novel meta-algorithm was highly successful in performance evaluations and **outperformed existing methods in 71% of trials.** The framework I developed can be applied to other problems such as dynamically managing resources harvested for serverless computing.

2. Predictions with unknown and variable noise levels

In addition to input predictors, system operators can leverage *parameter predictors*, which learn the correct value of a tunable parameter for an online algorithm. For example in the classical ski-rental problem, input predictions forecast whether a person will go skiing the next day, whereas parameter predictions forecast the total number of skiing days. Unlike input predictions, an online algorithm with access to perfect parameter predictions can enjoy the same performance as the offline optimal solution. This comes at the cost of exceptionally bad performance when the parameter prediction is of poor quality. An emerging area of research is utilizing *trust parameters* to exploit parameter predictions when they are of high quality while defaulting to conservative purely online decision making when the predictions have large error.

Trust parameters are a promising approach for cloud systems, where prediction quality can vary substantially due to shifting resource demands, as well as the complexity arising from interacting system-wide optimization, such as geographic load balancing and AI-powered thermal management. My collaborators and I had the first paper on using trust parameters in the data center setting, where we studied peak-aware energy generation scheduling of on-site generators [2]. Companies such as Microsoft reduce data center electricity bills by installing on-site generation units which can "shave the peak" and reduce the peak net demand from the grid over the billing cycle. We designed online deterministic and randomized algorithms that employ a user-determined trust parameter to determine how to optimally utilize on-site generation while balancing reliance on parameter predictions. The novel algorithms are **Pareto-optimal** with respect to (1) *robustness* (the competitive ratio as a function of the trust parameter when the prediction is inaccurate) and (2) *consistency* (the competitive ratio when the prediction is accurate). This work was a **best paper finalist at e-Energy '21.** The algorithms we developed can be extended to other rent-vs-buy problems such as server on/off scheduling, TCP acknowledgment and renting cloud servers.

3. Augmenting static strategies with dynamic predictions

A natural question is whether there is any benefit to using multiple types of predictors (ie. input and parameter). While parameter predictors offer a policy that mimics the offline static optimal, input predictors can be used to design a dynamic strategy over a much shorter time horizon. These two approaches fundamentally consider different time-scales and thus provide different advantages and disadvantages for resource allocation problems with switching costs. Further-

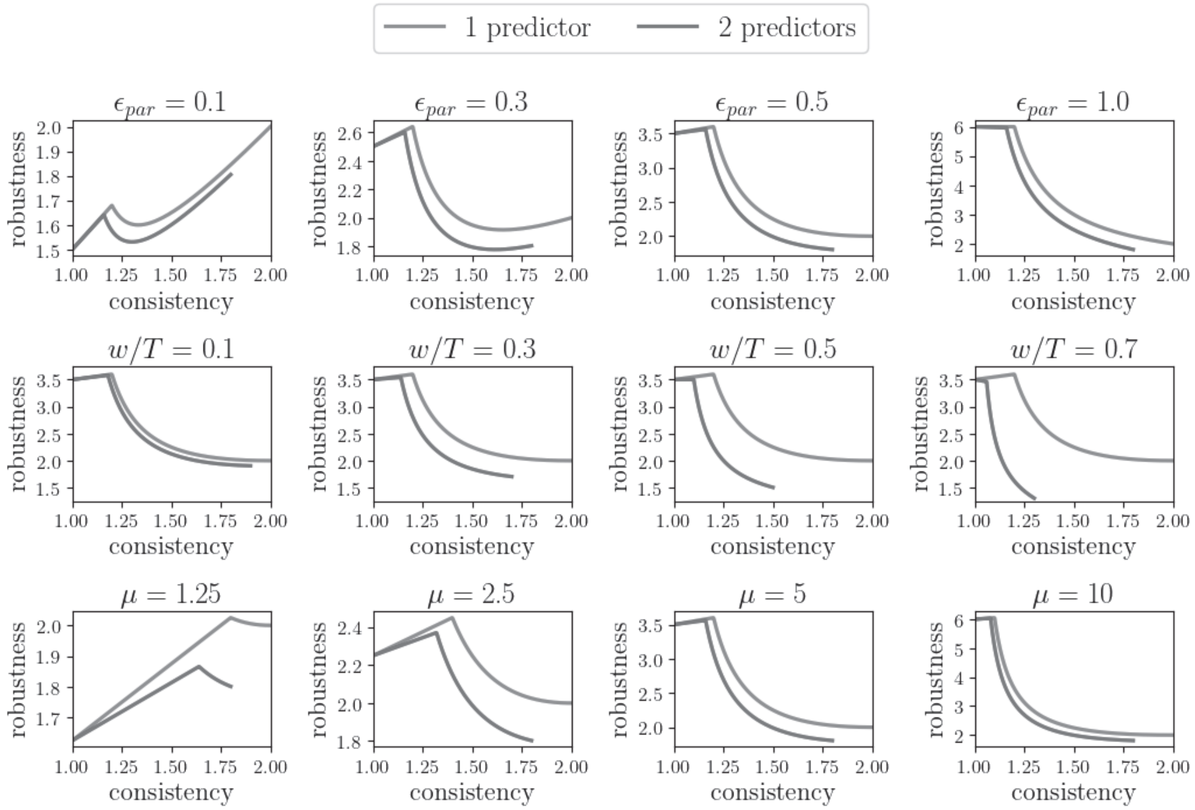

Figure 2: Visualizing the theoretical worst-case performance of using 1 vs 2 noisy predictors, for varying parameter prediction error ϵ_{par} (row 1), length of input prediction window w/T (row 2) and problem difficulty constant μ (row 3). In all plots, using 2 predictors consistently provides better performance than 1, for all values of trust parameter λ, despite the fact that the input predictions are imperfect and noisy.

more, there are many optimization problems where system operators have access to *heterogeneous* resources to satisfy user demands. These heterogeneous resources can vary in switching cost rates as well as operational cost rates, further adding to the time-coupling present in the optimization problem.

Despite the promise of using multiple predictor types, the current literature typically only considers online algorithms with input predictors. To this end, we designed an online algorithm that uses both predictions of the optimal static allocation as well as noisy forecasts of the future net user demand [4]. Our algorithm outperforms state-of-the-art competitors in worst-case performance guarantees by dynamically refining the policy suggested by the parameter prediction with updated forecasts of user demands in a short lookahead window. The performance of our algorithm was tested with extensive performance evaluations using real-world traces from the California Independent System Operator (CAISO). Our novel algorithm achieved a **17% improvement over the best competitor that only used input predictions** and a **92% improvement over the best online competitor**. These results illustrate the theoretical and empirical improvements achieved by leveraging access to good parameter predictions. The framework of using predictors that operate at different timescales is promising for many problems including chachine for efficient machine selection in VM scheduling.

4. Combining multiple noisy predictors

In practice, the quality of predictions are not known in advance nor whether assumptions about error levels will be satisfied, as assumed in the previous section. I tackled the challenge of different predictors of unknown quality in the context of bandwidth cost minimization for large distributed systems [3], [1]. Distributed services typically procure bandwidth from multiple data centers with different types of contracts and use an online algorithm to decide which contracts to use to satisfy time-varying and volatile demands, while minimizing the total bandwidth cost. The bandwidth cost minimization problem has a nice offline structure which is uniquely characterized by a break-even point. In practice, system operators will use parameter predictions of this break-even point to allocate demand among different contracts. However, the global COVID-19 pandemic has driven unpredictable internet traffic demands that resulted in a large discrepancy between predictions generated with historical data and current user demand.

For the input and parameter predictors, I designed online algorithms that can optimally utilize each predictor type and provided the first theoretical bounds for data-driven methods in the bandwidth cost minimization setting. The two predictors have complementary benefits of being effective in either the average-case or the worst-case, but not both. In the standard trust parameter framework, an algorithm would exploit high-quality predictions and default to classical online decision making when predictions are noisy. However, this work was motivated by the observation that even if predictions of one quantity are noisy, the classical online assumption of no information about the future might still be too harsh. And indeed, I proved that under very mild assumptions, using two noisy predictors outperforms only using one. This result holds for all settings of the trust parameter (see Figure 2), and even when the prediction noisy grows arbitrarily large. Our theoretical findings were empirically verified with performance evaluations using production traces from Akamai's content delivery network. This new framework demonstrates how to leverage diverse noisy predictors to achieve consistently strong performance and is a promising candidate for other problems such as power capping in cloud systems.

Future Work: The integration of multiple different predictors into online resource allocation algorithms has proven to be successful. System operators can leverage historical data to enjoy strong average-case performance, while still employing techniques from online algorithms to maintain worst-case performance guarantees. In the past year, I have become excited about the following two new directions: (1) combining distributional and instance-specific learning and (2) designing parsimonious online algorithms.

The projects I worked on for my thesis primarily focused on generating predictions for a given instance. However, historical data also can give insights into the entire distribution. An interesting open direction is whether it is possible to combine aggregate distributional information with predictions tailored to the instance at hand to achieve improved performance. One key technical challenge in this area is finding suitable low dimensional representations of distributional knowledge that can easily be leverage in an online manner.

The second direction addresses *parsimony*. Although ML predictions can be very insightful, they are costly to generate. Many current algorithms regenerate predictions of the same quantities, such as RHC and other popular control algorithms. An interesting direction is whether it is possible to maintain the same performance while significantly reducing the computational cost of generating predictions. In particular, a challenge open question is whether it is possible to maintain both parsimony and worst-case guarantees in non-stationary environments.

Representative Papers:

[1] Two Untrusted Predictors are Better than One: Minimizing Bandwidth Cost with Multiple Noisy Predictors (Under review)
with J. Maghakian, R. Lee, M. Hajiesmaili, J. Li, Z. Liu, R. Sitaraman

[2] Online peak-aware energy scheduling with untrusted advice (e-Energy 2021)
with R. Lee, J. Maghakian, M. Hajiesmaili, J. Li, Z. Liu, R. Sitaraman

[3] Leveraging Different Types of Predictors for Online Optimization (CISS 2021)
with J. Maghakian, R. Lee, M. Hajiesmaili, J. Li, Z. Liu, R. Sitaraman

[4] Online Economic Dispatch with Volatile Renewable Generation and Ramping Costs (ICNC 2020)
with J. Comden, J. Maghakian, Z. Liu

[5] Online Optimization in the Non-Stationary Cloud: Change Point Detection for Resource Provisioning (CISS 2019)
with J. Maghakian, J. Comden, Z. Liu