Proceedings of the ASME 2019
Dynamic Systems and Control Conference
DSCC2019
October 8-11, 2019, Park City, Utah, USA

DSCC2019-9068

EMPIRICAL REGRET BOUNDS FOR CONTROL IN SPATIOTEMPORALLY VARYING ENVIRONMENTS: A CASE STUDY IN AIRBORNE WIND ENERGY

Ben Haydon[1], Jack Cole[1], Laurel Dunn[2], Patrick Keyantuo[2], Tina Chow[2], Scott Moura[2], and Chris Vermillion[1]*

[1]Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC
[2]Department of Civil and Environmental Engineering, University of California - Berkeley, Berkeley, CA
*Please address all correspondence to this author.

ABSTRACT

This paper focuses on the empirical derivation of regret bounds for mobile systems that can vary their locations within a spatiotemporally varying environment in order to maximize performance. In particular, the paper focuses on an airborne wind energy system, where the replacement of towers with tethers and a lifting body allows the system to adjust its altitude continuously, with the goal of operating at the altitude that maximizes net power production. While prior publications have proposed control strategies for this problem, often with favorable results based on simulations that use real wind data, they lack any theoretical or statistical performance guarantees. In the present work, we make use of a very large synthetic data set, identified through parameters from real wind data, to derive probabilistic bounds on the difference between optimal and actual performance, termed regret. The results are presented for a variety of control strategies, including a maximum probability of improvement, upper confidence bound, greedy, and constant altitude approaches.

INTRODUCTION

Airborne Wind Energy (AWE) systems, such as Altaeros' Buoyant Airborne Turbine (BAT) [1], replace conventional towers with tethers and a *lifting body* (wing or aerostat) that holds the airborne generator aloft. These systems are of particular interest due to their ability to reach high altitudes, where wind speeds can be much higher than wind speeds typically seen by ground mounted turbines. Whereas a typical wind turbine is constructed and left in place, an AWE system has the capability of adjusting its altitude to seek the optimal conditions for power generation. The opportunities presented by control over altitude result in an important tradeoff. Because the wind speed is an unknown variable at all altitudes besides the current altitude of the AWE system, the controller must balance exploring other altitudes to search for better wind conditions (for more power generation later) and exploiting the current estimate of the maximum wind speed (for more power generation now). This problem is classified as exploration vs. exploitation in a spatiotemporally varying environment.

FIGURE 1: Altaeros Buoyant Airborne Turbine (BAT) [1]

A large body of research has examined different control

strategies for balancing exploration and exploitation in the context of AWE systems (see [2], [3], [4], [5], and [6]). Each of these results has demonstrated, based on simulations driven by real wind speed vs. altitude data, that appropriately designed altitude optimization algorithms can significantly outperform constant-altitude flight, even at the best constant altitude (which is only obtained through omniscient knowledge). The exploration/exploitation tradeoff addressed in the AWE literature has also been examined in wheeled mobile robotics literature (see [7], [8]), in addition to the unmanned aerial vehicle literature (see [9]). In these instances, the objective is to utilize a team of mobile robots to perform a mission in an unknown, spatiotemporally varying environment.

An examination of the aforementioned results reveals that while numerous strategies have been employed in the control of mobile systems in spatiotemporally varying environments, often very successfully, these results are not accompanied by theoretical or statistical performance guarantees. It is intuitive to focus on the difference between optimal and realized performance, termed *regret*, in formulating such bounds. In fact, a number of papers investigate the formulation of regret bounds in classical bandit problems [10, 11, 12]. Here, a gambler must, at every instance, select one of many arms to pull on, where each arm yields a different reward that is subject to some stochastic distribution. In the context of our problem, the gambler is the mobile (AWE) system, the arms are spatial locations (altitudes), and the rewards are performance outputs (power output or wind speed). Additionally, a few papers investigate the use of Gaussian Process modeling to determine regret bounds in bandit problems [13,14]. However, none of these papers provide a hard probabilistic bound on regret (e.g., with probability ε, regret will fall below a value of δ). Instead, the papers provide regret bounds that are specified in "big O" notation as a function of time, thereby identifying the manner by which regret will vary in time but not providing a hard bound. For practical purposes, this does not allow for calculable bounds that can be used to compare the relative strengths of different control strategies.

Focusing on AWE systems, this paper provides an empirical quantification of regret, calculated through a large synthetic data set that is informed by real wind speed vs. altitude data. In particular, we use wind speed vs. altitude data to identify statistical *hyperparameters* of a Gaussian Process (GP) model, then use this model to generate a synthetic wind data set. For several altitude control strategies, including maximum probability of improvement (MPI), upper confidence bound (UCB), greedy, and baseline constant altitude strategies, we compute statistical regret bounds using a Monte Carlo simulation setup. In particular, we generate empirical cumulative distribution functions (CDFs) for regret under each control strategy. This represents the first rigorous effort at "scoring" these spatiotemporal optimal control strategies.

The paper is organized as follows. First, we introduce the concept of regret and an important axiom that underlies our derivation of a CDF for regret under each control strategy. We then describe the process by which we generate a very large synthetic data set from existing wind data, followed by a description of altitude control strategies that are evaluated in the paper. We finish by presenting regret bound CDFs for each of the altitude control strategy.

MATHEMATICAL PRELIMINARIES

For the control strategies evaluated in this paper, the variable $z(t)$ will be used to represent the spatial decision variable (the chosen altitude in the AWE application), $P(z(t))$ will be used to represent the reward (power generation) at the chosen z, for a given time, t.

When judging control strategies operating in a partially observable environment, it is important to utilize a quantitative performance metric. In this paper, we work with a quantity known as *regret*. Put simply, regret is the difference in performance between the optimal strategy given perfect knowledge and the strategy chosen by the controller based on imperfect knowledge. Mathematically:

$$r(t) = \mathbb{E}\left[P^*(z(t)) - P(z(t)\right], \tag{1}$$

where $r(t)$ is the instantaneous regret, or the regret at a single instant in time, $P^*(z(t))$ is the current maximum value of the performance function over the entire spatial domain, and $P(z(t)$ is the value of the performance function evaluated at the point chosen by the control strategy. In many instances, cumulative regret and average regret are more informative statistics than instantaneous regret. These quantities are given by:

$$R(t) = \sum_{\bar{t}=1}^{t} r(\bar{t}) = \sum_{\bar{t}=1}^{t} E(P^*(z(\bar{t})) - P(z(\bar{t}))), \tag{2}$$

$$R_{avg}(t) = \frac{R(t)}{t}, \tag{3}$$

where $R(t)$ represents cumulative regret and $R_{avg}(t)$ represents average regret. These metrics are especially useful because maximizing overall performance over a certain time interval is often more important than attempting to maximize performance at each point in time. In fact, the best strategy will necessarily sacrifice some of its current performance in order to explore the payoffs at other locations. This is necessary so that the system can determine whether any other point has become more profitable. This trade-off is often referred to as exploration vs. exploitation.

Having defined regret, we now turn to the quantification of regret bounds. Because the spatiotemporally varying environment is stochastic in nature, there always exists a chance that any control strategy will yield zero reward at any time. Consequently, it is impossible to derive an upper bound on regret (other than a trivial zero-reward bound) with 100% confidence. To circumvent this issue, we examine regret bounds in a *probabilistic* sense. In particular, for every regret bound, there will exist some confidence level with which that regret bound can be achieved. Equivalently, given a specified confidence bound, there will exist some corresponding regret bound that will be achieved with that level of confidence. This relationship is encoded in the following axiom:

Axiom For every $\varepsilon \in (0,1), T > 0$, $\exists\ \delta(\varepsilon) > 0$ such that

$$Pr[r(t) \leq \delta(\varepsilon)] = 1 - \varepsilon, \forall t \geq T \tag{4}$$

This relationship can also be expressed in terms of average regret:

$$Pr\left[R_{avg}(t) \leq \bar{\delta}(\bar{\varepsilon})\right] = 1 - \bar{\varepsilon}, \quad \forall\, t \geq T \tag{5}$$

The key contribution of this paper lies in the estimation of $\delta(\varepsilon)$ (and $\bar{\delta}(\bar{\varepsilon})$ in the case of average regret) for different control strategies; the regret bound acts as a scoring mechanism for each strategy. It is worth noting that the relationship between δ and $1-\varepsilon$ is equivalent to the cumulative distribution function (CDF) for regret. Two sample relationships between δ and ε are illustrated in Figure 2. By definition, ε is bounded between 0 and 1. Given that low regret is desirable, a $\delta - \varepsilon$ curve closer to the ε-axis describes a superior control strategy. If two curves cross, neither strategy is superior in an absolute sense.

DATA-DRIVEN SYNTHETIC WIND MODEL

To use empirical methods to derive regret bounds of the form of Equations (4) and (5), a large data set is required. In order to obtain this, actual wind speed vs. altitude data from [15] was first used to extract critical statistical properties that describe the temporal and spatial evolution of the wind speed. This data has been obtained by a 900 MHz Doppler wind profiler deployed in Lewes, Delaware. Data is available for nearly the entirety of a year, at altitudes up to 3km.

The characteristics identified from the aforementioned training data serve as the backbone for the generation of a much larger *synthetic* data set, which is based on a Gaussian Process (GP) model, as described below.

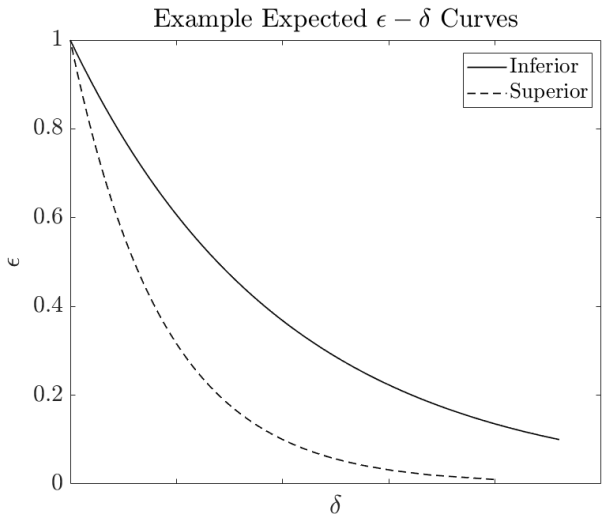

FIGURE 2: Expected Shape of $\varepsilon - \delta$ Curve.

Overview of Model

A GP model can be used to describe a stochastic process for which a collection of random variables is described via a mean function and covariance function. Available MATLAB tools for GP modeling [16] were used in conjunction with the aforementioned wind data to estimate the mean function and tune the hyperparameters of the covariance function, namely the spatial and temporal length scales.

The mean function of a GP model defines the predicted mean at any location if there is no data available upon which to condition a prediction. We used a fourth-order polynomial, $m(z)$, to define the mean function for the GP, as shown in Equation (6). The coefficients of this mean function are its hyperparameters. It is key to note that the assumed mean function does not vary in time, meaning that simulated data can be generated for any length of time. The parameterization of the mean function is given by:

$$m(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4 \tag{6}$$

Based on the data that were used in identifying the mean function, we fit the coefficients a_0, a_1, a_2, a_3, and a_4 to the data used in identifying the mean function with least squares regression. The polynomial coefficients are given by:

$$a_0 = 13.366\frac{m}{s}, \quad a_1 = 9.103\frac{m}{s \cdot km},$$
$$a_2 = 2.596\frac{m}{s \cdot km^2}, \quad a_3 = -2.707\frac{m}{s \cdot km^3}, \quad a_4 = 0.535\frac{m}{s \cdot km^4}$$

The covariance of a GP model defines the amount of corre-

lation between two points. This level of correlation is a function of the distance and time between the two points. The parametric function used by the GP model to encode this correlation is referred to as a *covariance kernel*. In this work, we use a squared exponential covariance kernel, which has the form:

$$K(z,t,z',t') = s^2 e^{-\frac{z-z'}{2l_z^2}} e^{-\frac{t-t'}{2l_t^2}}. \tag{7}$$

Here, s, l_z, and l_t are the *hyperparameters* of the kernel, which are identified from data. Specifically, s^2 is the signal variance, which characterizes the expected deviation (squared) of a given wind measurement from the mean. The variables l_z and l_t are the length scales of the system, and they characterize how quickly the wind profile changes with respect to altitude and time, respectively. Based on the calibration of the model, l_t and l_z are given by:

$$l_t = 22 \text{ min}, \quad l_z = .27 \text{ km}$$

The aforementioned mean and covariance functions are related to two very important quantities, namely the *prediction mean* and *prediction variance*. These are the conditional mean and variance of the prediction error, conditioned upon data collected up until that point in time, and are given by [16]:

$$\mu_* = m(x_*) + K(x_*,x)K(x,x)^{-1}(\mathbf{y} - m(x)) \tag{8}$$

$$\sigma_*^2 = K(x_*,x_*) - K(x_*,x)K(x,x)^{-1}K(x,x_*) \tag{9}$$

where x_* is a test point for which the prediction mean μ_* and prediction variance σ_*^2 are calculated. The functions $m(\cdot)$ and $K(\cdot,\cdot)$ are the mean and covariance functions defined in Equations (6) and (7). The vector $\mathbf{y}$ is the set of training outputs corresponding to the inputs x. In this paper, each x is a 2-element vector containing altitude (z) and time (t). The prediction mean and prediction variance are critical in defining the control strategies that balance exploration and exploitation. Furthermore, they are also critical in generating synthetic data.

The method for generating the GP model consisted of two main steps. First, the model's hyperparameters were calibrated based on the available wind data, using the methodology from [16]. The hyperparameters were chosen to maximize the log marginal likelihood over 100 evaluations of the model with respect to the training data, which is calculated as follows:

$$-\log(p(\mathbf{y}|x)) = \frac{1}{2}\mathbf{y}^T K(x,x)^{-1}\mathbf{y} + \frac{1}{2}log|K(x,x)| + \frac{n}{2}\log 2\pi \tag{10}$$

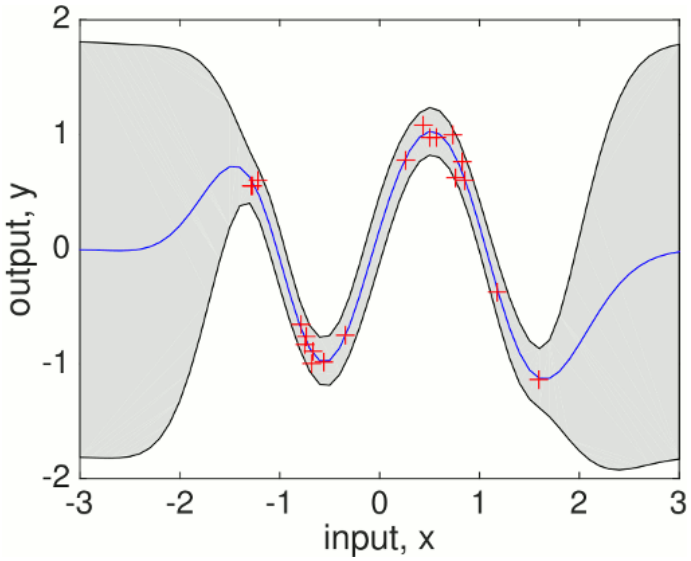

FIGURE 3: 2-D Example GP Model Mean $\pm 2\sigma$ [16]

where n is the number of training points. Using these hyperparameters, the GP model calculates the prediction mean and variance of any point, given a set of observed data. A 2-D example of this is shown in Figure 3. The points represent observed data, the blue curve represents the prediction mean, and the gray region represents two standard deviations (square root of the prediction variance) in either direction of the mean. Note that at positions far from any observed data, the prediction variance is large due to the fact that there is no data in the vicinity upon which to calculate the prediction mean.

Generating Synthetic Data

Given parameters of the GP model, synthetic data is generated by marching forward in time, using synthetic data up to that point in time to characterize prediction mean and variance at the next time step, then generating random data at that next time step based on that mean and variance. To initialize the model, an initial wind profile, $v_w(t_0,z)$, was chosen. To generate $v_w(t_1,z)$, $v_w(t_0,z)$ was fed into the GP model, which was used to compute $\mu_t(z)$ and $\sigma_t(z)$ at t_1. The wind speed at each altitude, at time t_1, was then computed as a Gaussian random variable with mean $\mu_t(z)$ and standard deviation $\sigma_t(z)$. This synthetic wind profile was then appended to the synthetic data, and the synthetic data for time t_2 (and all future times) were generated by applying the same process for each time step.

On-Board Estimates for Control

Ultimately, each candidate control strategy must make a decision (which altitude to operate at) based on available data. In all candidate strategies, the chosen altitude is a function of the

prediction mean and prediction variance, *based on the data at altitudes that have been visited up to that time*, which represents a subset of the total set of synthetic data. This is where the GP model shines; it effectively allows for the calculation of a prediction mean and variance based upon any amount of available data.

Figure 4 shows a sample set of synthetic data in the top left plot, along with three sets of on-board estimated wind profiles, each based on a different control strategy (where the details of the control strategies are given in the subsequent section). Figure 5 shows each control trajectory overlaid on a contour plot of synthetic data. Because the data made available under each control strategy comprises only a subset of the available synthetic data, deviations between the on-board estimates and synthetic data are observed and indeed expected. Furthermore, because each control strategy results in the exploration of a different sequence of altitudes, variation in the on-board models is observed and indeed expected.

CONTROL STRATEGIES

In this section, we compare five control strategies for altitude control: Maximum Probability of Improvement (MPI), Upper Confidence Bound (UCB), a "greedy" algorithm, and two constant altitude strategies, defined as "constant altitude - omniscient" and "constant altitude - average." It is important to note that the two constant altitude strategies require omniscient knowledge; thus, they are not real-time implementable strategies but rather merely serve as benchmarks against which the other altitude control strategies can be compared.

Each control strategy operates based on the maximization of an *acquisition function*, $\alpha_t(\mu_t(z),\sigma_t(z))$, which is always a function of the prediction mean (μ_t) and prediction variance (σ_t^2). In particular, each control law takes the form:

$$z(t) = \arg\max_{\bar{z}\in Z} \alpha_t(\mu_t(\bar{z}),\sigma_t(\bar{z})), \tag{11}$$

where Z represents the domain of allowable control variables (altitudes for the AWE system). By tailoring the structure of $\alpha(\mu_t(z),\sigma_t(z))$, it is possible to obtain different tradeoffs between exploration and exploitation.

Upper Confidence Bound

The Upper Confidence Bound (UCB) [17] control strategy explicitly trades off exploration and exploitation through the following acquisition function:

$$\alpha_t(\mu_t(z),\sigma_t(z)) = \mu_t(z) + \sqrt{\beta_t} * \sigma_t(z), \tag{12}$$

where β_t is a parameter that defines the relative weighting of $\mu_t(z)$ and $\sigma_t(z)$ in the acquisition function. To determine an acceptable value of β_t, several simulations were run with different β_t values, and $\beta_t = 1$ was determined to be near-optimal. This value of β_t means that the acquisition function will be defined as the prediction mean plus one standard deviation. A more optimistic algorithm would have a higher β_t value, giving a higher weighting to the *possibility* of improvements past one standard deviation.

Each time the acquisition function is evaluated, points with high variance and points with high expected value are both valued. This is an example of "optimism in the face of uncertainty." Although the environment is uncertain, the algorithm is optimistic about its chances if it visits an altitude with a high potential wind speed, even if the expected wind speed at that altitude is low. The outcome of this strategy is that at least one of two results will arise from each control action: Either the performance (wind speed) will be high or the system will learn a significant amount about a point in the spatial domain that was previously poorly characterized.

Maximum Probability of Improvement (MPI)

The MPI strategy [18] manages the exploration/exploitation tradeoff in a different manner than the UCB approach. As the name suggests, instead of a linear function of mean and standard deviation, the MPI acquisition function is simply the probability that the reward for visiting some altitude is greater than the greatest reward seen so far. The corresponding acquisition function is given by:

$$\alpha_t(\mu_t(z),\sigma_t(z)) = \Phi\left(\frac{\mu_t(z) - P_{max}}{\sigma_t(z)}\right), \tag{13}$$

where Φ is the cumulative distribution function with a normal distribution and P_{max} is the greatest reward seen so far.

"Greedy" Algorithm

Slivkins and Upfal [10] define a "greedy" algorithm that simply selects the control value with the highest expected reward, then remains there for a selected amount of time. Its acquisition function is therefore given by:

$$\alpha_t(\mu_t(z),\sigma_t(z)) = \mu_t(z). \tag{14}$$

The number of time steps for which the strategy decides to remain at each chosen altitude is given by:

$$\tau = l_t\sqrt{\log l_t} - k, \tag{15}$$

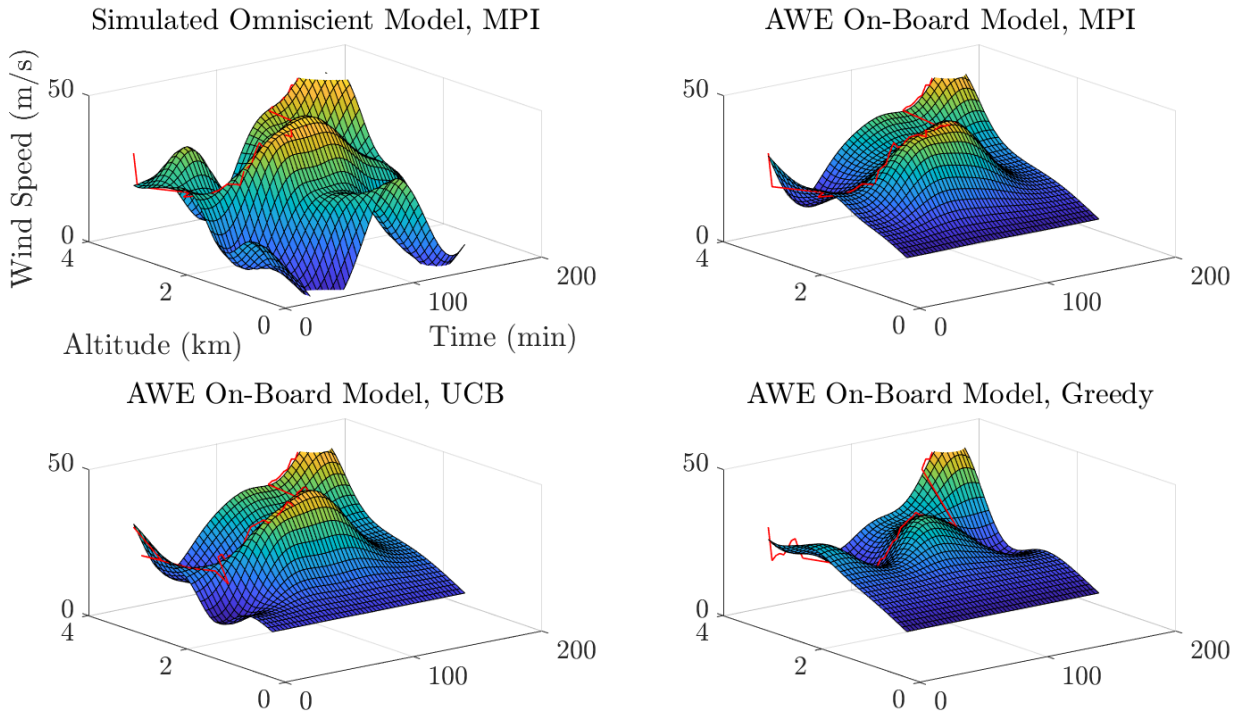

FIGURE 4: Comparison of Several On-Board Models to Synthetic Generated Data

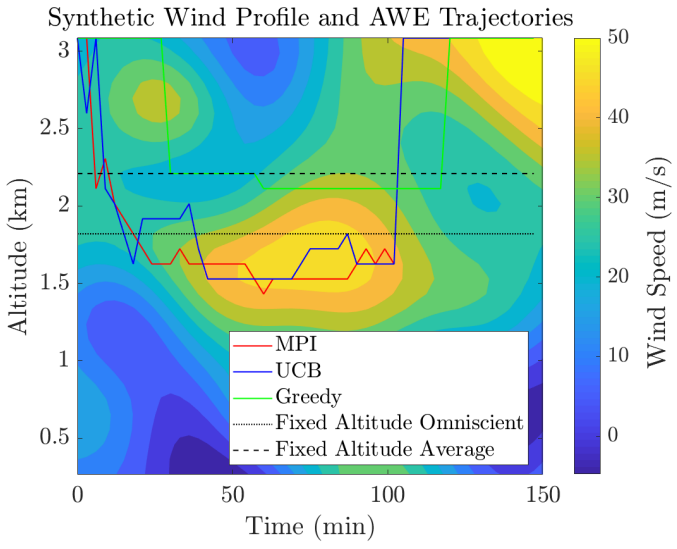

FIGURE 5: Contour Plot showing AWE Trajectories

Where l_t is the time scale of the environment, and k is the number of discretized altitudes to choose from. This algorithm is termed greedy because it ignores exploration in its acquisition function. However, the amount of time spent at each altitude is a function of time scale to account for variability of the system.

Fixed Altitude - Omniscient (FAO)

The fixed altitude omniscient strategy simply involves staying at the constant altitude that provides the highest average reward. It is important to note that this strategy is not real-time implementable, as it requires omniscient knowledge of the wind data over the entire time window and all altitudes. This strategy merely represents a benchmark against which variable altitude algorithms can be compared.

Fixed Altitude - Average (FAA)

The fixed altitude average algorithm involves remaining at the altitude that results in an average reward equal to the mean of all possible rewards. Like the constant altitude omniscient strategy, this strategy is not real-time implementable but instead represents a benchmark against which variable altitude algorithms can be compared.

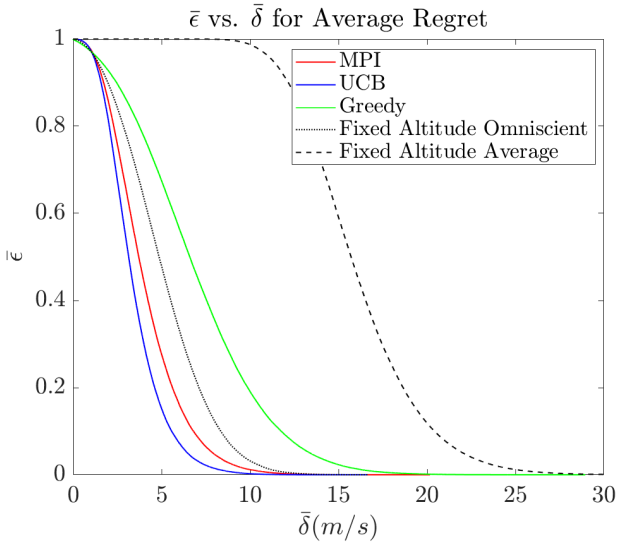

FIGURE 6: Average Regret Epsilon-Delta Plot

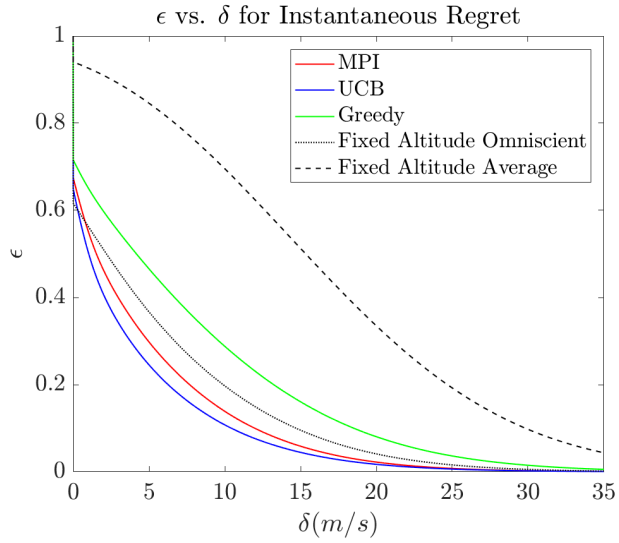

FIGURE 7: Instantaneous Regret Epsilon-Delta Plot

RESULTS

Using each of the acquisition functions defined above, 100,000 simulations were run, each with a different set of synthetic wind data generated as described earlier in the paper. Instantaneous and average regret were logged for each run of each control strategy. For every simulation, each control strategy operated independently under the same set of synthetic data. These synthetic data were used to generate the regret bound comparisons of Figs. 6 and 7, which describe the relationship between δ and ε, as defined in Equations (4) and (5).

Several conclusions can be drawn from these results. First, each algorithm outperformed the average stationary strategy by a significant margin, so each has some merit. However, the greedy strategy failed to outperform the omniscient stationary strategy, reinforcing the idea that the explicit incentivization of exploration in the acquisition function is important to long-term exploitation. Both UCB and MPI outperformed the omniscient stationary strategy by a significant margin, showing that each is an effective algorithm in a wide range of cases.

Control Strategy	MPI	UCB	Greedy	FAO	FAA
$\bar{\delta}(m/s)$	7.90	6.47	13.39	9.38	22.03

TABLE 1: Comparison of Average Regret Upper Bounds with 95% Confidence

A useful property of these graphs is the ability to compare individual points as well. For instance, comparing each strategy based on a 95% confidence interval on average regret can be done by comparing the $\bar{\delta}$ values at $\bar{\varepsilon} = .05$. The resulting average regret bounds at this 95 percent confidence level are given in Table 1.

Another key fact to note is that both plots seem to have an asymptote along the δ axis. This result is intuitive, since there is no way to provide a hard upper bound on regret. It is interesting that each graph has a different trend as it approaches $\varepsilon = 1$. Because it is possible to obtain zero instantaneous regret (in a discretized spatial environment), there is a range of ε over which δ is zero under all algorithms. Comparing this to the average regret plots, $\bar{\delta}$ increases extremely quickly in going from $\bar{\varepsilon} = 1$ to $\bar{\varepsilon} = .99$. This also makes sense because it is nearly impossible to obtain zero cumulative (and therefore average) regret.

CONCLUSIONS AND FUTURE WORK

Focusing on AWE systems as a case study for optimal control in a spatiotemporally varying environment, this paper presented the formulation of a Gaussian Process model for use in generating synthetic wind speed profiles varying in time. These wind models were then used to compare Upper Confidence Bound (UCB), Maximum Probability of Improvement (MPI), and greedy algorithms in terms of instantaneous and cumulative regret. UCB outperformed MPI, which outperformed the greedy algorithm by a significant margin.

Future work will focus on expanding the characterization to include other exploration/exploitation strategies, particularly extremum seeking and model predictive control strategies developed in [2] and [5], which both explicitly incorporate exploration and exploitation into their objective functions. Further work will also investigate the dependence of regret bounds on spatiotemporal length scales, along with an investigation into analytical expressions for regret bounds and a comparison with the empiri-

cal results presented herein.

ACKNOWLEDGMENT

This work was supported by NSF Award Number 1711579, entitled "Collaborative Research: Multi-Scale, Multi-Rate Spatiotemporal Optimal Control with Application to Airborne Wind Energy Systems."

REFERENCES

[1] Yallop, O., 2014. "Inflatable turbines: the windfarms of the future?". Accessed: 2019-03-29.

[2] Bafandeh, A., and Vermillion, C., 2016. "Real-time altitude optimization of airborne wind energy systems using lyapunov-based switched extremum seeking control". *American Control Conference*, 07, pp. 4990–4995.

[3] Baheri, A., Bin-Karim, S., Bafandeh, A., and Vermillion, C., 2017. "Real-time control using bayesian optimization: A case study in airborne wind energy systems". *Control Engineering Practice,* **69**, 12, pp. 131–140.

[4] Bafandeh, A., Bin-Karim, S., Baheri, A., and Vermillion, C., 2018. "A comparative assessment of hierarchical control structures for spatiotemporally-varying systems, with application to airborne wind energy". *Control Engineering Practice,* **74**, 12, pp. 71–83.

[5] Bin-Karim, S., Bafandeh, A., Baheri, A., and Vermillion, C., 2017. "Spatiotemporal optimization through Gaussian Process based model predictive control: Case study in airborne wind energy". *IEEE Transactions on Control Systems Technology,* **27**, 3, pp. 798–805.

[6] Dunn, L., Vermillion, C., Chow, F. K., and Moura, S., 2019. "On wind speed sensor configurations and altitude control in airborne wind energy systems". *American Control Conference (accepted)*.

[7] Solanas, A., and Garcia, M., 2004. "Coordinated multi-robot exploration through unsupervised clustering of unknown space". *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[8] Quann, M., Ojeda, L., Smith, W., Rizzo, D., Castanier, M., and Barton, K., 2017. "An energy-efficient method for multi-robot reconnaissance in an unknown environment". *American Control Conference*.

[9] Bentz, W., Hoang, T., Bayasgalan, E., and Panagou, D., 2018. "Complete 3-d dynamic coverage in energy-constrained multi-uav sensor networks". *Autonomous Robots,* **42**, pp. 825–851.

[10] Slivkins, A., and Upfal, E., 2008. "Adapting to a changing environment: the brownian restless bandits". In 21st Conference on Learning Theory (COLT), pp. 343–354.

[11] Besbes, O., Gur, Y., and Zeevi, A., 2014. "Stochastic multi-armed-bandit problem with non-stationary rewards". In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds. Curran Associates, Inc., pp. 199–207.

[12] Garivier, A., and Moulines, E., 2011. "On upper-confidence bound policies for switching bandit problems". In Algorithmic Learning Theory, J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, eds., Springer Berlin Heidelberg, pp. 174–188.

[13] Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W., 2009. "Gaussian Process bandits without regret: An experimental design approach". *CoRR*.

[14] Krause, A., and Ong, C. S., 2011. "Contextual Gaussian Process bandit optimization". In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds. Curran Associates, Inc., pp. 2447–2455.

[15] Archer, C., 2014. Wind Profiler at Cape Henlopen.

[16] Rasmussen, C. E., and Nickisch, H., 2018. GPML Matlab Code version 4.2.

[17] Cox, D., and John, S., 1992. "SDO: A statistical method for global optimization". *Multidisciplinary Design Optimization: State of the Art*, 11, pp. 1241 – 1246 vol.2.

[18] Kushner, H. J., 1964. "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise". *Journal of Basic Engineering*, pp. 97 – 106.