RESEARCH STATEMENT

Gautam Goel

The goal of my research is to combine machine learning with signal processing and control to obtain novel data-driven algorithms for inference and control in dynamical systems. This area is currently undergoing a renaissance due to the broad range of potential applications, which include control of robotic systems such as drones and autonomous vehicles; analysis of biological signals, such as EKGs and EEGs; decentralized control of cyberphysical systems, such as the smart grid; and climate modeling and weather forecasting. My work focuses on answering two complementary questions:

Q1: How do we incorporate learning and adaptivity into signal processing and control? Traditional algorithms for signal processing and control in dynamical systems posit a specific model for the environment encountered by the algorithm. For example, the Kalman filter is derived under the assumption that the observations fed into the algorithm are perturbed by stochastic noise which is zero-mean and i.i.d.; this assumption may be too optimistic in practice. Similarly, robust controllers assume that the environment is adversarial; this assumption is often too pessimistic. A natural idea is to use ideas from machine learning to design adaptive algorithms, which dynamically adjust to their environment to achieve optimal performance across many different settings. These algorithms should achieve the Best of Both Worlds: in stochastic settings they should perform as well as algorithms specifically designed for stochastic environments, but they should also retain strong performance guarantees even in adversarial environments. In a similar vein, it is usually assumed that the algorithm has a perfect model of the system dynamics. In practice, this is rarely the case; there are always some aspects of the system which will remain unmodeled. How do we design algorithms which improve their own internal model of the dynamics over time to make better decisions?

Q2: How do we extend online learning to settings with non-stationarity and dynamics? Online learning is a subfield of machine learning which studies how an intelligent agent should interact with an external environment so as to maximize the rewards it collects. Classical online learning algorithms are designed to perform well in *stationary* environments, where the rewards associated to actions do not change over time. In such an environment, the reward-maximizing strategy is simply to learn which single action is associated with the highest reward and then to select that action in every round. If, however, the environment is non-stationary, the reward-maximizing strategy is to identify a time-varying sequence of actions, with actions being selected more frequently when they are associated with high rewards. Classical online learning theory also assumes that the rewards are not coupled across rounds, i.e. the actions of the learner do not affect the rewards available in the future. In applications such as control neither assumption holds; the rewards associated to each action vary as the external environment changes, and the actions selected by the algorithm directly affect the state of the system, which in turn affects future rewards. The application of online learning in such applications hence calls for different performance metrics and new approaches to algorithm design.

In a series of joint works with several collaborators, I have developed a new theoretical framework to address these questions. This framework combines recent ideas from online learning with classical tools from H_∞ estimation and control, resulting in novel data-driven algorithms with strong empirical performance; these algorithms have both learning-theoretic and control-theoretic interpretations. I briefly summarize these results, and describe how I would like to extend my research program to reinforcement learning, statistical learning theory, and nonlinear estimation and control.

Regret-Optimal Filtering and Control

Motivation

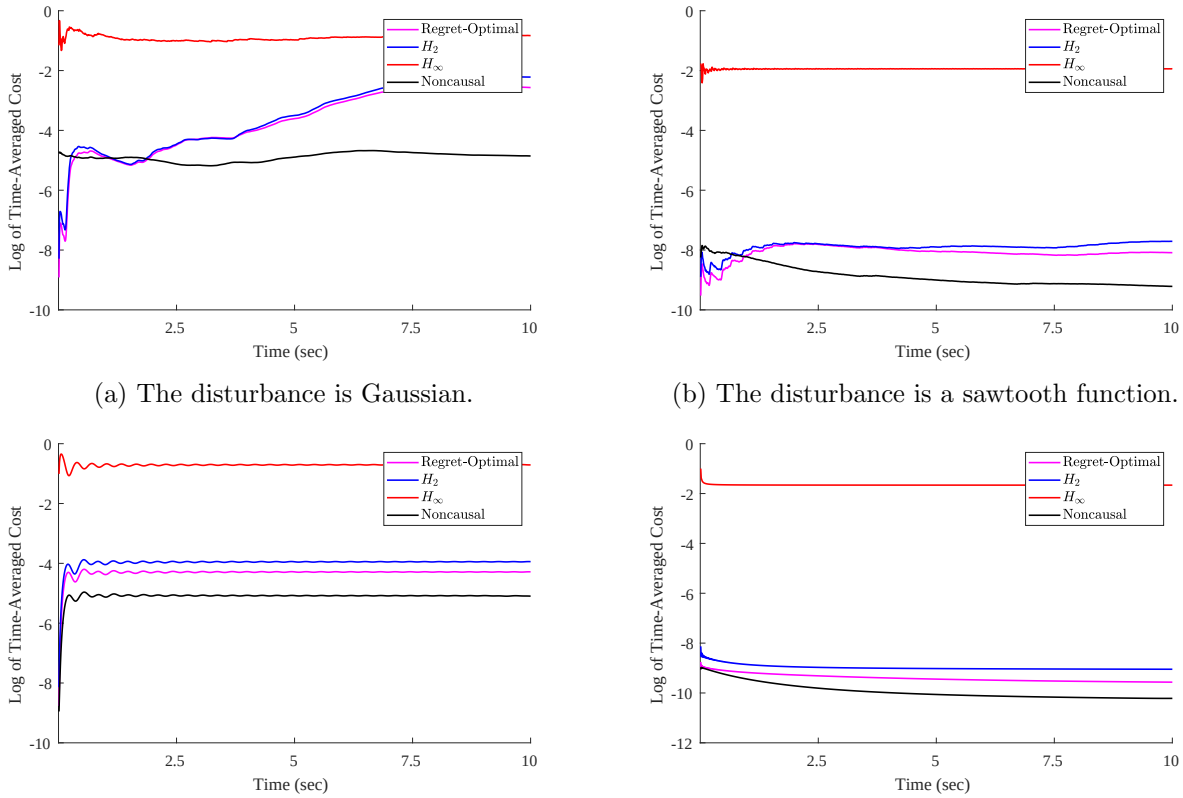
Optimal control frames control as an optimization problem, where the goal of the controller is to minimize a cost function which measures how far away the system is from the desired trajectory and how much energy the controller is expending. In many real-world settings, the controller is unable to directly observe the state but must instead estimate the state indirectly given noisy measurements of the system. The problem of constructing this estimate is called *state estimation* or *filtering*. This problem has been widely studied in the signal processing community, leading to well-known algorithms like the Kalman filter and the particle filter.

Filtering and control have traditionally been studied in two distinct settings. In the H_2 setting, the disturbances and measurement noise are both assumed to be generated independently by zero-mean stochastic processes, and the goal of the controller is to minimize its expected cost. The Kalman filter is the H_2-optimal filter and the well-known LQG controller is the H_2-optimal measurement-feedback controller. In the H_∞ setting, control is formulated as a minimax game played between the controller and an adversary which selects the disturbances and measurement noise; the goal of the controller is to minimize its own cost, while the goal of the adversary is to force the controller to incur high cost. Intuitively, this formulation leads to "robust" controllers with bounded worst-case performance.

Both of these design methodologies suffer from an obvious drawback: a controller which encounters a disturbance which falls outside of the class it was designed to handle may perform poorly. Indeed, experience shows that H_∞ controllers tend to be over-conservative, whereas H_2 controllers have poor performance outside of stochastic settings. This observation naturally motivates the design of adaptive controllers which dynamically adjust their control strategy as they observe the disturbance instead of blindly following a prescribed strategy. Inspired by online learning, I have explored how *regret* can be used as a design criterion for filters and controllers. Regret is the standard performance metric in online learning; it is the difference between the cumulative rewards collected by the online algorithm and the rewards which could counterfactually have been selected given the benefit of hindsight. Intuitively, a learning algorithm with low regret is one which is resilient to uncertainty - it performs almost as well as a hypothetical algorithm which could make the optimal decisions with perfect knowledge of the relationship between actions and rewards.

My contributions

In Regret-Optimal control, we adopt a very different perspective from classical H_2 and H_∞ control: instead of explicitly modeling the disturbances and designing the controller accordingly, we instead design a controller whose cost is as close as possible to the cost which could be obtained by a hypothetical "noncausal" controller which knew the disturbance sequence in advance and selected the cost-minimizing sequence of actions accordingly. Similarly, in Regret-Optimal filtering, we design a filter whose cost is as close as possible to the cost which could be obtained by a hypothetical noncausal filter which received the whole sequence of measurements at once and selected the optimal (in a least-squares sense) sequence of state estimates. Surprisingly, I showed that Regret-Optimal controllers can be derived using a reduction to H_∞ control [1, 2, 4, 5]; a controller with optimal regret in a given linear dynamical system can be viewed as an H_∞ controller in a synthetic system of larger dimension (a similar reduction allows one to construct Regret-Optimal filters [3, 4]). The extra states of the synthetic system can be viewed as a "sufficient statistic" that summarizes which information about the past the algorithm needs to store to improve its decisions in the future.



(a) The disturbance is Gaussian.

(b) The disturbance is a sawtooth function.

(c) The disturbance is sinusoidal.

(d) The disturbance alternates between $+1$ and -1.

Figure 1: We plot the relative performance of the H_2, H_∞, competitive, and noncausal controllers in the nonlinear inverted pendulum system. The competitive controller consistently outperforms the H_2 controller by at least a factor of 2 and the H_∞ controller by orders of magnitude. The hypothetical noncausal controller outperforms all other controllers. Note that the y-axis is on a log scale.

The reduction from Regret-Optimal control to H_∞ control has several desirable properties: First, the resulting Regret-Optimal controller is the controller with *optimal* regret out of *all possible controllers*, including nonlinear controllers. Unlike prior work in learning and control, I do not restrict my attention to asymptotic regret bounds (e.g. $O(\log(T)), O(\sqrt{T})$, etc) or a specific parametric class of controllers, such as state-feedback policies. Second, the synthetic system can be efficiently computed given the parameters of the original system; the Regret-Optimal controller is also efficiently computable using standard H_∞ synthesis. Lastly, a Regret-Optimal controller exists whenever an H_2 or H_∞-optimal controller exists; the construction of Regret-Optimal controllers does not rely on any non-standard assumptions.

Although Regret-Optimal controllers are derived in the context of control in linear systems, they can still be applied to systems with nonlinear dynamics using Model Predictive Control (MPC). In the MPC framework, we iteratively linearize the model dynamics around the current state, compute the optimal control signal in the linearized system, and then update the state in the original nonlinear system using this control signal. Numerical experiments using MPC in the nonlinear inverted pendulum system (Figure 1) show that the competitive controller (i.e., the controller with optimal competitive ratio) outperforms standard H_2 and H_∞ controllers, often by orders of magnitude; this may be because Regret-Optimal controllers integrate learning and are hence better able to adapt to nonlinear dynamics.

Together, these results show an exciting and previously unsuspected interplay between online learning and robust estimation and control. While H_∞ theory was originally developed with the aim of ensuring *robustness*, my work shows that it can be repurposed to ensure *adaptivity*. In future work I would like to extend these ideas to richer settings with nonlinear or decentralized dynamics; recent work from ETH Zürich studies how to incorporate safety constraints into Regret-Optimal control [11].

Online Learning 2.0

Another major thrust of my research is to understand how online algorithms should be designed for learning in complex, time-varying environments, where the rewards obtained by the algorithm in each round do not depend solely on the decision made by the algorithm in that round, but potentially on all previous choices made by the algorithm.

In one line of work [6, 7], I studied online learning in non-stationary environments, specifically in the context of Online Convex Optimization (OCO) with switching costs. While most works in online learning study how an online algorithm can learn the best fixed action in a stationary environment, this problem studies how an online algorithm should adapt its decisions in the face of a changing environment. Furthermore, in this problem the online algorithm incurs a penalty every time it changes its action; the rewards obtained by the algorithm are hence coupled across rounds. This models many applications where it is expensive to change your decision, such as dynamic right-sizing in datacenters [9]. Resolving a decades-old open problem, I obtained the first online algorithm for this problem which attained a constant competitive ratio in metric spaces of arbitrary dimension; this algorithm is called Online Balanced Descent (OBD) since it balances the rewards it obtains from selecting good actions with the cost of switching actions. I also showed that OBD achieves bounded competitive ratio when applied to online control, thus establishing a surprising connection between OCO and control. While OBD does not match the optimal competitive ratio attainable with Regret-Optimal control, it does have one advantage: it is *memoryless*, meaning that it makes decisions using only the information currently available to it, and does not need to store the full history of observations. More recently, I've studied how predictions can be leveraged to improve the performance of online algorithms, even in settings where the rewards are nonconvex in the actions [10].

More recently, I have explored how we can combine Regret-Optimal control with online learning algorithms to obtain novel "hybrid" algorithms. In forthcoming work with several collaborators, I showed that Regret-Optimal controllers are closely approximated by a class of control policies called disturbance-action-control (DAC) policies. These policies have attracted much recent attention in the learning and control literature, because the convexity of this class means that it can be efficiently optimized online using standard gradient-descent based algorithms from online learning. In addition, it is known that regret bounds against this class also imply regret bounds against the better-known class of state-feedback controllers (which include H_2 and H_∞ controllers). These results imply that an algorithm which is initialized to select the same control actions as a Regret-Optimal controller but which also updates its internal parameters using online learning can simultaneously minimize regret against the optimal sequence of actions possible in hindsight (like Regret-Optimal algorithms) and against DAC and state-feedback policies (like gradient-descent based online learning algorithms). Furthermore, we can combine a Regret-Optimal controller with the recently proposed algorithm in [12] to obtain a controller with optimal competitive ratio (up to a sublinear correction) even in systems with unknown dynamics. I believe that this hybrid approach to controller design, which supplements an optimal controller with an iterative, gradient-based online learning layer, holds much promise in the area of learning and control; similar ideas could be used to construct adaptive filters.

Future Directions

Online Decision-Making with Complex Reward Structure. I am interested in understanding how statistical learning theory can be used to improve online decision-making. Online learning algorithms generally form an internal model of the relationship between actions and rewards; this model is continually improved using observations collected by the algorithm, and the updated model is in turn used to improve the decisions of the algorithm. Traditional online learning considers only very simple models; for example, a standard assumption in the contextual bandit problem is to assume that the expected reward is a linear function of the context. More recently, there has been much interest in using richer models (such as deep neural networks) to model the relationship between actions and rewards, e.g. [8, 16]. The advantage of this approach is that it allows for online learning in much richer settings, where the rewards are a complex, nonlinear function of the actions and contexts. Unfortunately, an understanding of how the performance of the online algorithm depends on the complexity of the model class remains poor.

Residual Learning in Nonlinear Systems. While Regret-Optimal algorithms have demonstrated favorable empirical performance in systems with nonlinear dynamics, they retain no provable guarantees in such settings. One approach to machine learning in nonlinear systems which I would like to explore is *residual learning*. In residual learning, we first approximate the true dynamics by a simple model (often a linear model), and compute the cost-minimizing controller in this approximate model. We then use deep learning to obtain a control policy which corrects for the "residual", i.e. the error in the approximate model. Intuitively, this combines the strengths of classical optimal control and modern machine learning; a similar idea could be used to construct a nonlinear filter. While residual learning works very well in practice [13], it currently lacks a rigorous theory to explain its empirical successes. A key technical challenge is to understand how the performance of the policy learned with a residual network scales with the amount of training data (i.e., sample complexity).

Best of Both Worlds in Reinforcement Learning. Reinforcement learning (RL) and optimal control are intimately related; the goal of both fields is to determine how an agent should steer the trajectory of an evolving system so as to maximize its cumulative rewards. While RL has historically focused on settings where the state and action spaces are finite (tabular RL), over the past decade there has been a surge in interest in settings where the state and action spaces are exponentially large (as in games such as Go or StarCraft) or continuous (as in robotics and control). Unfortunately, the high dimensionality of these spaces means that it is computationally infeasible to use standard tabular RL techniques to learn an optimal policy; state-of-the-art RL systems hence resort to heuristics like function approximation and deep Q-learning, which lack a sound theoretical justification. At the same time, there is growing interest in moving beyond worst-case analysis of RL to obtain algorithms with optimal instance-dependent performance, i.e. algorithms which retain worst-case guarantees but can take advantage of problem structure to perform better on easier instances (this Best of Both Worlds philosophy also motivates Regret-Optimal control). Unfortunately, prior work (e.g. [14, 15]) is restricted to the tabular setting. A major open problem is thus to obtain instance-optimal RL algorithms for high-dimensional, continuous action spaces.

Shannon meets Hamming: Best of Both Worlds in Coding Theory. A somewhat speculative direction for future work is to study how online learning can be integrated into coding theory. Coding theory is often studied under two error models; a stochastic model (the Shannon model) and an adversarial model (the Hamming model). Consider two parties sequentially transmitting data to each other over a noisy channel - is there a transmission protocol that quickly converges to the optimal rate under both models? In other words, is there a protocol which achieves the Best of Both Worlds - efficient data transfer in both stochastic and adversarial settings?

References

[1] Gautam Goel and Babak Hassibi. "Competitive Control". In: *arXiv preprint arXiv:2107.13657* (2021).

[2] Gautam Goel and Babak Hassibi. "Measurement-Feedback Control with Optimal Data-Dependent Regret". In: *arXiv preprint arXiv:2209.06425* (2022).

[3] Gautam Goel and Babak Hassibi. "Online estimation and control with optimal pathlength regret". In: *arXiv preprint arXiv:2110.12544* (2021).

[4] Gautam Goel and Babak Hassibi. "Regret-optimal Estimation and Control". In: *arXiv preprint arXiv:2106.12097* (2021).

[5] Gautam Goel and Babak Hassibi. "Regret-optimal measurement-feedback control". In: *Learning for Dynamics and Control*. PMLR. 2021, pp. 1270–1280.

[6] Gautam Goel and Adam Wierman. "An online algorithm for smoothed regression and lqr control". In: *Proceedings of Machine Learning Research* 89 (2019), pp. 2504–2513.

[7] Gautam Goel et al. "Beyond online balanced descent: An optimal algorithm for smoothed online optimization". In: *Advances in Neural Information Processing Systems*. 2019, pp. 1875–1885.

[8] Parnian Kassraie and Andreas Krause. "Neural contextual bandits without regret". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 240–278.

[9] Minghong Lin et al. "Dynamic right-sizing for power-proportional data centers". In: *IEEE/ACM Transactions on Networking* 21.5 (2012), pp. 1378–1391.

[10] Yiheng Lin, Gautam Goel, and Adam Wierman. "Online optimization with predictions and non-convex losses". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 4.1 (2020), pp. 1–32.

[11] Andrea Martin et al. "Safe control with minimal regret". In: *Learning for Dynamics and Control Conference*. PMLR. 2022, pp. 726–738.

[12] Edgar Minasyan et al. "Online control of unknown time-varying dynamical systems". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15934–15945.

[13] Guanya Shi et al. "Neural lander: Stable drone landing control using learned dynamics". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 9784–9790.

[14] Max Simchowitz and Kevin G Jamieson. "Non-asymptotic gap-dependent regret bounds for tabular mdps". In: *Advances in Neural Information Processing Systems* 32 (2019).

[15] Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. "Beyond no regret: Instance-dependent pac reinforcement learning". In: *Conference on Learning Theory*. PMLR. 2022, pp. 358–418.

[16] Dongruo Zhou, Lihong Li, and Quanquan Gu. "Neural contextual bandits with ucb-based exploration". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 11492–11502.