Provably Optimal Control/Learning for Model Problems

PIs: Charlie Fefferman, Clancy Rowley

Grad students: Daniel Gurevich
               Amlan Sinha
               Melanie Weber

Postdocs: Jacob Carruth
          Max Eggl
          Debdipata Goswami
Outline

Overview of the model problems

Complete results for model problem 1 and several variants

Partial results for problem 2 and variants

Initial progress on variants of problem 3
The Model Problems

- Apply a time-dependent control $u$ to keep the position $q$ of a moving particle close to zero.
- $u(t), q(t) \in \mathbb{R}$.
- Dynamics of $q$ depend on $u$, random noise and one unknown parameter $a$.
- The particle starts at time zero at $q = 0$.
- We are free to apply any control we please, but we’re not allowed to look into the future.
Dynamics of the moving particle

Three models:

TM I: \( dq = (a + u)dt + dW \)

TM II: \( dq = (aq + u)dt + dW \)

TM III: \( dq = au \, dt + dW \)

where

\( q = \text{position}, \quad u = \text{control}, \quad t = \text{time}, \)

\( dW = \text{White noise}, \)

\( a = \text{unknown parameter} \)
We want to apply a control strategy $u$ to minimize the expected value of

$$J = \int_0^T \left( q^2 + \lambda u^2 \right) dt,$$

where $\lambda > 0$ is a known coefficient.

- If $a$ is known, that’s a standard control theory problem.
- If $a$ is unknown, then as time $t$ increases, we learn more and more about $a$ from history up to time $t$. So we are doing control with learning on the fly.
Our approach vs. other approaches

Our approach differs from existing approaches in several respects:

▶ Some standard approaches (e.g., reinforcement learning) require lots of data, from many trials. We do not have that luxury.

▶ Other approaches define success as scaling well with the time horizon $T$ in the limit as $T \to \infty$.

▶ Our approach achieves provably optimal performance for a fixed time horizon $T$, given only observations from a single trial.
Our model problems vs. real problems

- Our model problems are much simpler than problems encountered in real life.
- However, our model problems already force us to face significant issues. We hope to “climb a ladder” by solving ever harder problems, starting with our model problems.
We now present a few variants of the model problems:

- Problem 1a: vector-valued position and control, with sensor noise
- Problem 2a and 3a: periodic spatial domain
- Problem 2b and 3b: restricted set of allowed values of parameter $a$
Problem 1a: vector-valued position and control

- Dynamics $dq = (a + Bu)dt + dW$
  - Both $q$ and $u$ are vector-valued
  - $a$ is an unknown drift vector
  - $B$ is a known matrix
  - $dW$ is a multivariate Gaussian with known covariance

- Cost $J = \int_0^T (q^T Q q + u^T R u) \, dt$
  - $Q$, $R$ known symmetric positive-definite matrices
Problem 1a (continued): sensor noise

- We are no longer allowed to observe the position \( q(t) \).
- Instead we observe \( y(t) = q(t) + \text{noise} \)
- The noise is given by a mean-reverting Brownian motion
In this variant, \( q(t) \) is an angle, so changing it by \( 2\pi \) has no effect

Problem 2a
- Dynamics: \( dq = (u + a \sin q) dt + dW \)
- Cost: \( J = \int_0^T [(1 - \cos q) + u^2] \, dt \)

Problem 3a
- Dynamics same as problem 3: \( dq = au \, dt + dW \)
- Cost \( J = \int_0^T [(1 - \cos q) + u^2] \, dt \)
In model problems 2, 2a, 3, and 3a, we might decide that the unknown parameter $a$ is one of the following:

- an unknown member of a given finite set
- an unknown number in a given interval
- a completely arbitrary unknown real number
How to formulate the problem?

Bayesian control vs. Agnostic control

- Bayesian: assume a probability distribution reflecting our prior belief about the unknown $a$.

- Pick our strategy to minimize the expected value of $J$. 
Agnostic control

What to do if we know **absolutely nothing** about the unknown $a$? (no prior belief)

How to formulate the problem?
The idea of regret

- We play against an opponent who knows the value of $a$.
- We pick our control strategy.
- Opponent picks optimal strategy for $a$.

\begin{align*}
J_{\text{our team}}(a) &= \text{Expected value of } J \text{ obtained using our strategy, given the value of } a. \\
J_{\text{opponent}}(a) &= \text{Expected value of } J \text{ obtained by opponent, given the value of } a.
\end{align*}
The idea of regret

Additive regret

\[ AR(a) = J_{\text{our team}}(a) - J_{\text{opponent}}(a) \geq 0 \]

Multiplicative regret (a.k.a. competitive ratio)

\[ MR(a) = \frac{J_{\text{our team}}(a)}{J_{\text{opponent}}(a)} \geq 1 \]

These depend on our strategy.
If the cost is small, multiplicative regret isn’t useful.

If the cost is large, additive regret isn’t useful.

This motivates a “hybrid regret” which we define as follows:

Fix a constant $c \geq 0$

Find a strategy such that, for any $a$,

$J_{our\,team}(a) \leq \gamma(J_{opponent}(a) + c)$ with $\gamma$ as small as possible
Agnostic control problems

Fix our favorite notion of regret.

**Problem:** Find a strategy such that for any $a$

$$\text{Regret}(a) \leq \gamma$$

with constant $\gamma$ as small as possible.
Outline

Overview of the model problems

Complete results for model problem 1 and several variants

Partial results for problem 2 and variants

Initial progress on variants of problem 3
For the first model and several variants

\[ dq = (a + u)dt + dW \]

- We know strategies with optimal additive or multiplicative regret.
- Both are Bayesian strategies with regret independent of the unknown parameter \( a \).
- The optimal strategies do surprisingly well: e.g., for multiplicative regret we can take \( \gamma \leq 1.17 \).
Even for the worst-case value of $T$, the optimal strategy for multiplicative regret is only 17% worse than the optimal strategy for known $a$. 
Variants of problem 1

We have similar results for

- the vector version of problem 1
- the version of problem 1 with sensor noise
- the vector version of problem 1 with sensor noise
Outline

Overview of the model problems

Complete results for model problem 1 and several variants

Partial results for problem 2 and variants

Initial progress on variants of problem 3
Recall model problem 2

- Dynamics given by

\[ dq = (aq + u)dt + dW \]

- Minimize a cost

\[ J = \int_0^T \left( q^2 + \lambda u^2 \right) dt \]

- The parameter \( a \) could be
  - Arbitrary
  - Confined to an interval
  - Required to belong to a finite set \( A \)
Solving the Bayesian problem

- For a given prior belief about $a$, the optimal control strategy is determined by the solution of a Bellman equation.
- That Bellman equation is a PDE in three space variables and time.
- In some cases we are able to solve this equation numerically.
We look for Bayesian strategies for a prior belief with the following properties:

- The regret is independent of $a$ as $a$ varies in a set $A_{\text{active}}$.
- Our prior probability distribution is concentrated entirely on $A_{\text{active}}$.
- For all $a$ outside $A_{\text{active}}$, the regret is less than its constant value on $A_{\text{active}}$. 
Optimal agnostic control

- If we can find such a prior probability distribution, then the resulting Bayesian strategy is optimal for agnostic control.
- We have also a converse result, but the exact statement is technical.
A significant difference between problems 1 and 2

- In problem 1, every $a$ belongs to the active set $A_{\text{active}}$
- For problem 2, the active set $A_{\text{active}}$ is finite
A good example

▶ \( A_{\text{active}} = \{-1, 1\}; \) prior belief makes regret equal for these two values of \( a \).

▶ For \( a \in (-1, 1) \), the regret is less than this constant value.

▶ Therefore, the resulting Bayesian strategy is optimal for multiplicative regret for all \( a \) in the interval \( A = [-1, 1] \).
A bad example

\[ A_{\text{active}} = \{-7.5, 1\} \]; prior makes regret constant on this set.

Now, the regret for intermediate values of \( a \) is not always less than this constant value.

So the resulting Bayesian strategy is not optimal for \( a \in [-7.5, 1] \).
Another good example

- $A_{\text{active}} = \{-5.227, 1\}$; prior makes regret constant on this set.
- The regret is less than this constant value for $a$ in the interval $A = [-12.57, 1]$
- So the resulting Bayesian strategy is optimal for all $a$ in $A$. 
For hybrid regret with parameters $c = 0.02$ and $T \leq 12$, we identify an optimal strategy for $a \leq 1$.

Depending on $T$, our prior belief for the optimal strategy is concentrated on two or three points.

It appears that for larger $T$ the regret is smaller than for the above cases.
Hybrid regret for $T = 10$

For $T = 10$, a three-point strategy minimizes $\gamma$
Agnostic control performs surprisingly well: for all values of $T \leq 12$, we have $\gamma < 1.17$. 

![Graph showing the relationship between optimal $\gamma$ and final time $T$.]
Comparison with other approaches

We compare with other control strategies:
- Optimal control for known $a$
- Optimal control assuming a parameter value that best fits previous observations (linear regression)
- Online gradient descent (E. Hazan et al., 2020)
Comparison with other approaches ($T = 10$)

- Known a
- Agnostic
- Linear regression
- Online gradient descent

Cost $J$ vs. $a$
We can solve Bayesian problems.

The optimal agnostic control is a Bayesian strategy for a special prior belief.

Unlike model problem 1, this special prior belief may be concentrated on finitely many points (e.g., only two, three, or four points).

We have solved the problem if the maximum allowed $a$ is not too big.

If the maximum allowed $a$ is too big, numerical issues arise. We have to overcome them.
Outline

Overview of the model problems

Complete results for model problem 1 and several variants

Partial results for problem 2 and variants

Initial progress on variants of problem 3
In this problem $a \in \mathbb{R}$ is an unknown parameter, $u \in \mathbb{R}$ is our control, $dW$ is white noise, and the position $q$ evolves according to

$$dq = (au)dt + dW.$$ 

We would like to choose a strategy $u$ to minimize the cost

$$J = \int_0^T (q^2 + \lambda u^2) dt$$

where $\lambda > 0$ is a parameter and $T > 0$ is the final time.
We work with a periodic version of problem 3.

To simplify the problem, let us assume $a$ is either $+1$ or $-1$.

We observe two regimes for the Bayesian problem:

- A “tame” regime, in which our control $u$ stays bounded
- A “wild” regime, in which our control $u$ is unbounded
The two regimes for problem 3

The “tame” regime
In the “tame” regime, the optimal strategy comes from a PDE: the Bellman equation.

The “wild” regime
In the “wild” regime, $u$ is large, so we make “big kicks”.

- More precisely, we take $u$ large until we gather enough information to place us back in the “tame” regime.
- Because $u$ is large, this happens almost immediately.
What is a strategy?

- In principle, the least possible regret arises as the limit of such strategies as the $u$ we use in the “wild” regime tends to infinity.
- In practice, we then want to use the largest attainable $u$.
- So the definition of a strategy should be broadened to allow infinite controls $u$ for infinitesimally short times $\Delta t$.
- An interesting problem to consider in the future is what happens when we impose constraints on the maximum allowable $u$. 
These model problems are much simpler than real-world problems, but already each problem has illuminated several features of optimal agnostic control:

► **TP1:** Optimal strategies are Bayesian with constant regret
► **TP2:** Optimal strategies are Bayesian with constant regret only on a subset
► **TP3:** There are regimes in which there is pure exploration: the wild regime
Future prospects

- We want to incorporate sensor noise into model problems 2 and 3
- We want to learn optimal strategies for problems in which both the dynamics and the effect of the controls are unknown (combining problems 2 and 3).
- For higher-dimensional versions of model problems 2 and 3, it is unlikely that we will be able to solve PDEs in many spatial dimensions, so we will need to find ways to produce computationally feasible upper and lower bounds for the optimal regret.
- Such lower bounds will be useful in evaluating heuristic (non-optimal) strategies for control-oriented learning on the fly.