Model-based exploration tuning

Jesse Clifton, Lili Wu, Eric Laber

North Carolina State University

February 25, 2019
Outline

Introduction

Review of bandits and MDPs

Review of exploration strategies

The parameterized exploration meta-heuristic

Simulations

Conclusion
Section 1

Introduction
Running decision problem example

Controlling glucose levels

- Maahs et al. (2012) conducted mHealth study of patients with type 1 diabetes
- Data collected using glucose monitor (CGM), accelerometer, digital insulin pump logs, and food logs
- Covariates observed at time $t$ include blood glucose level $Gl^t$, dietary intake $Di^t$, and physical activity $Ex^t$
- Actions: Whether to recommend insulin, activity, and/or food at each time point
- Goal: Keep glucose levels in normal range
The exploration-exploitation tradeoff

- When making a sequence of decisions under uncertainty, one must trade off:
  1. Acting based on current beliefs about the best thing to do (exploitation)
  2. Acting in order to gain more accurate beliefs

- We focus on cases where the goal is to maximize cumulative rewards (corresponding to, say, patient outcomes)

- However, the discussion may be adapted to cases where exploration is also of primary interest (e.g., adaptive allocation in sequential experiments)
  - Villar et al. (2015) use finite-horizon bandit approach to adaptive clinical trial design
The need for careful exploration

- Common heuristics for exploration in reinforcement learning
  - Are designed for infinite-horizon problems
  - Disregard knowledge of system dynamics due to intractability in short-decision-interval settings
- On the other hand, in some applications
  - Time horizons are short
  - Statistical efficiency takes precedence over computational efficiency (so can use more expensive exploration procedures)
- We present a simple family of algorithms for tuning the amount of exploration in sequential decision-problems
- Our procedure leverages knowledge of the underlying generative model, as well as the time horizon of the decision problem
Section 2

Review of bandits and MDPs
Sequential decision-making

Sequential decision-making problems are characterized by

- A reward function
- Set of available actions at each step
- Environment governed by unknown probability distributions
- Goal is to maximize some measure of cumulative expected reward over the course of the decision problem

In more complicated models, reward distributions also depend on states (e.g. contextual bandits) and may be dependent over time (e.g. Markov decision processes)

Generic notation

- Time horizon $T < \infty$
- Decision points indexed by $t = 1, \ldots, T$
- Actions $A^t$ in action space $\mathcal{A}$
- Rewards $U^t$
- Environment characterized by generative model $\mathcal{M}$ (will see concrete examples)
Multi-armed bandits

- Actions are “arms” $\mathcal{A} = \{1, \ldots, k\}$
- Reward distributions $\mathcal{M} = (D_1, \ldots, D_k)$ for each action
  - So $U^t \sim D_{A^t}$
  - Denote reward means as $\mu_1^M, \ldots, \mu_k^M$ and $\mu^{*M} = \max_i \mu_i^M$.
- Denote observed sequences of actions and rewards up to time $t$ as $A^t = (A^1, \ldots, A^t)$ and $U^t = (U^1, \ldots, U^t)$, and thus the history $H^t = (A^t, U^t)$
Multi-armed bandits

- Goal is to take a sequence of actions $A^T$ which maximizes expected cumulative reward

$$V(A^T, M) \triangleq \sum_{t=1}^{T} \mu_{M_{At}}$$

- Since we don’t know $M$, we have to trade off learning about $M$ (exploring) and acting according to our current estimates of each $\mu_i^M$ (exploiting)
Define a exploration strategy $\Gamma = (\Gamma^1, \ldots, \Gamma^T)$, where for each $t$ $\Gamma^t$ maps histories $H^t$ to probability distributions over actions

- Not the same as policies $\pi$ in dependent decision processes
- A exploration strategy may involve estimating a policy!

Exploration strategies usually consist of an estimator of the optimal strategy + an exploration heuristic

e.g. $\epsilon$-greedy

$$\Gamma^t \equiv \Gamma_{\epsilon^t} : H^t \mapsto \begin{cases} \arg \max_i U^t_i, & \text{with probability } 1 - \frac{\epsilon^t}{k}, \\ j, & \text{with probability } \frac{\epsilon^t}{k}, j = 1, \ldots, k; \end{cases}$$

where $U^t_i$ the sample mean of rewards from action $i$
Multi-armed bandits

- The optimal exploration strategy over a class \( \mathcal{G} \) is

\[
\Gamma^* \triangleq \arg \max_{\Gamma \in \mathcal{G}} \mathbb{E} \sum_{t=1}^{T} \mu_{\Gamma}^{\mathcal{M}}(H_t)
= \arg \max_{\Gamma \in \mathcal{G}} \mathbb{E}_{A \sim (\Gamma, \mathcal{M})} V(A^T, \mathcal{M}).
\]

- Let \( C \) be a distribution encoding uncertainty over \( \mathcal{M} \)
  - e.g. Estimated sampling distribution, posterior
- Then the “\( C \)-optimal” exploration strategy in \( \mathcal{G} \) is

\[
\Gamma^*_C \triangleq \arg \min_{\Gamma \in \mathcal{G}} \mathbb{E}_{\mathcal{M} \sim C} \left[ \mathbb{E}_{A \sim (\Gamma, \mathcal{M})} V(A^T, \mathcal{M}) \right].
\]
Markov decision processes

- State space $S$; at each time step take action $A^t$ in state $S^t$
- State evolves according to Markovian transition probabilities
  - $\mathcal{M}$ characterized by transition function $P(S^{t+1} \mid S^t, A^t)$
- Reward function depends on states as well as actions; $U^t \equiv u(S^t, A^t)$
Example MDP: Controlling glucose levels

- A continuous-state MDP based on the mHealth study of Maahs et al. (2012) of patients with type 1 diabetes
- Data collected using glucose monitor (CGM), accelerometer, digital insulin pump logs, and food logs
- Covariates observed at time $t$ include blood glucose level $Gl^t$, dietary intake $Di^t$, and physical activity $Ex^t$
Example MDP: Controlling glucose levels

- Actions are give insulin or not, $A^t \in \{0, 1\}$
- Writing $x^t \leftarrow (Gl^t, Di^t, Ex^t, A^t)$, glucose levels are modeled as AR(2) process:

\[
Gl^{t+1} = \beta_0 + \beta_1^T x^t + \beta_2^T x^{t-1} + e^{t+1},
\]
\[
e^{t+1} \sim N(0, \sigma^2)
\]

- Food and activity modeled as iid at each time point
- Reward function $u$ is a function which increases as glucose departs from normal levels
- This is an MDP with states $s^t \leftarrow (x^t, x^{t-1})$
Markov decision processes

- A policy is a sequence of decision rules $\pi = (\pi_t)_{t=1}^T$ mapping states to probability distributions over actions, i.e. $\pi_t : S \to \Delta(A)$

- The policy-value function $V$ measures the cumulative expected reward under each policy

$$V(\cdot, M) : \pi \mapsto \mathbb{E}_{(S^t, A^t) \sim (\pi, M)} \left[ \sum_{t=1}^{T} u(S^t, A^t) \right].$$

- The optimal policy is defined as

$$\pi^{\text{opt}} \triangleq \arg \max_{\pi} V(\pi, M).$$
Markov decision processes

- If the time horizon is infinite, there is a stationary optimal policy (i.e. $\pi_t$ is constant)
- Exploration strategies in the MDP setting usually consist in
  1. Treating time horizon as infinite
  2. Estimating optimal infinite-horizon policy $\hat{\pi}^t$
  3. Trading off taking $\hat{\pi}^t(S^t)$ and using exploration heuristic
Section 3

Review of exploration strategies
Review of exploration strategies

- We de-emphasize theoretical analysis of a given exploration strategy’s regret
- View the exploration strategy as a tuning parameter
- Focus on strategies for MABs; easily ported to MDPs
Review of exploration strategies

$\epsilon$-greedy

- Take greedy action with high probability, act uniformly at random otherwise
- Let $\overline{U}^t_a$ be the sample mean of rewards from arm $a$ at time $t$
- Then for a sequence of values $\epsilon^t \in [0, 1]$, the $\epsilon$-greedy exploration strategy is

$$
\Gamma_{\epsilon^t} : H^t \mapsto \begin{cases} 
\arg \max_a \overline{U}^t_a, \text{ with probability } 1 - \frac{\epsilon^t}{k}, \\
j, \text{ with probability } \frac{\epsilon^t}{k}, j = 1, \ldots, k;
\end{cases}
$$
Upper confidence bound (UCB)

- Act according to optimistic estimates of the value of each arm
- Let $\mathcal{U}^t(a, \alpha)$ be a $1 - \alpha$ confidence bound on $\mu^*_a$ at time $t$
- e.g. $\mathcal{U}^t(a, \alpha) \equiv \overline{U}_a^t + z_{1-\alpha} SE_a^t$, where $SE_a^t$ is standard error of rewards from arm $a$ at time $t$

$$\Gamma_{\alpha^t} : H^t \mapsto \arg \max_a \mathcal{U}^t(a, \alpha^t).$$
Review of exploration strategies

Thompson sampling

- Let $P(\mu_a \mid \mathbf{H}^t)$ be a posterior over value of action $a$ given history at time $t$
- For each, sample $\tilde{\mu}_a^t \sim P(\mu_a \mid \mathbf{H}^t)$
- Take $A^t \leftarrow \arg\max_a \tilde{\mu}_a^t$
Review of exploration strategies

Modified Thompson sampling

- Let $P(\mu_a \mid \mathbf{H}^t; \tau)$ be a probability distribution over $\mu_a$ whose variability is controlled by parameter $\tau$
- Examples
  - If $P(\mu_a \mid \mathbf{H}^t)$ has scale-family density $f(\cdot \mid \mathbf{H}^t)$, $P(\mu_a \mid \mathbf{H}^t; \tau)$ has density $\tau f(\cdot/\tau \mid \mathbf{H}^t)$
  - Chapelle and Li (2011) use modified Beta-Binomial posterior with parameters $(\alpha/\tau, \beta/\tau)$, observe improved small-sample performance
- The modified Thompson sampling exploration strategy is

$$\Gamma_{\tau^t} : \mathbf{H}^t \mapsto \arg \max_{a} \left\{ \tilde{\mu}^t_a \sim P(\mu_a \mid \mathbf{H}^t; \tau^t) \right\}$$
A few other strategies

- **Bayes-optimal**: Follow the policy of maximizing posterior expected cumulative reward (generally intractable)
- **Curiosity / intrinsic motivation**: Modify the reward function to give a bonus for learning about the environment
- **Entropy regularization**: Modify the objective for estimating an optimal policy to encourage higher-entropy policies (i.e. those which spread probability mass among actions more)
Section 4

The parameterized exploration meta-heuristic
Parameterized exploration strategies

- Let $\{\eta^t\}_{t=1}^T$ be a sequence of parameters of a exploration strategy, e.g. $\eta^t = \epsilon_t$ in $\epsilon$-greedy
- Parameterize $\eta^t$ by a family of functions increasing in the remaining time $T - t$,

$$\eta^t \equiv \eta(T - t, \theta)$$

for some $\theta \in \Theta$

- Each value of $\theta$ leads to an exploration strategy

$$\Gamma_{\theta} \triangleq (\Gamma_{\eta(T-1,\theta)}, \ldots, \Gamma_{\eta(T,T,\theta)})$$

(suppressing dependence on $T, \eta$)

- Write $\mathcal{G}_\Theta \triangleq \{\Gamma_{\theta} : \theta \in \Theta\}$
Parameterized exploration schedule

- Example class of decay schedules:

\[
\eta(T - t, \theta) \triangleq \frac{\theta_0}{1 + \exp[-\theta_2(T - t - \theta_1)]}
\]

Figure 1: Example estimated optimal decay schedules for \(\epsilon\) in a Gaussian multi-armed bandit.
Parameterized exploration strategies

- Let $\hat{M}^t$ be an estimator of the generative model $M$ at time $t$
  - For MABs, $M = D$
  - For Markov decision processes, $M = \text{Transition function}$
- Define (for MABs)

$$V^T(\theta, M) \triangleq \mathbb{E}_M \left[ \sum_{t=1}^{T} \mu_{\Gamma_{\eta(T-t, \theta)}(H^t)}^{\text{M}} \right]$$

the expected cumulative regret incurred by $\Gamma_{\theta}$ when $M$ is the true model.
Parameterized exploration meta-heuristic

- A plugin estimator of the optimal exploration strategy in $G_\Theta$ is

$\hat\theta^t \triangleq \arg\min_{\theta \in \Theta} V^T(\theta, \hat{M}^t)$.

- For uncertainty distribution $C^t \equiv C(H^t)$, the $C^t$-optimal exploration strategy in $G_\Theta$ is

$\hat{\theta}^t_C \triangleq \arg\max_{\theta \in \Theta} \mathbb{E}_{\hat{M} \sim C^t} V^T(\theta, \hat{M})$.

- $\hat{\theta}^t_C$ is preferable to $\hat{\theta}^t$, as $\hat{\theta}^t$ will tend to under-explore (corroborated by preliminary simulation results)
Parameterized exploration meta-heuristic

- This suggests a meta-heuristic for exploration: At each timestep, estimate $\hat{M}^t$ and act according to $\Gamma_{\hat{\theta}^t_C}$

---

**Algorithm 1** Parameterized exploration for MABs

**Input** Function class $\{\eta(\cdot, \cdot, \theta) : \theta \in \Theta\}$; decision rule class $\{\Gamma_\eta : \eta \geq 0\}$; reward distributions $\{D_i\}_{i=1}^k$; time horizon $T$

- $A^0 = \{1, \ldots, k\}$
- $U^0 = \{U^0_1 \sim D_1, \ldots, U^0_k \sim D_k\}$
- $H^0 = \{A^0, U^0\}$

**for** $t = 0, \ldots, T$ **do**

- Obtain $C^t$ from $H^t$
- $\hat{\theta}^t_C \leftarrow \arg \max_{\theta \in \Theta} \mathbb{E}_{\tilde{M} \sim C^t} V^T(\theta, \tilde{M})$
- $A^t \sim \Gamma_{\eta(T, t, \hat{\theta}^t_C)}(H^t)$
- $U^t \sim D_{A^t}$
- $H^{t+1} \leftarrow H^t \cup \{A^t, U^t\}$
Section 5

Simulations
In each of our simulations, we used the class
\[
\eta(T - t, \theta) = \frac{\theta_0}{1 + \exp[-\theta_2(T - t - \theta_1)]} \mid (\theta_0, \theta_1, \theta_2) \in \Theta
\]

\[\mathbb{E}_{\mathcal{M} \sim C_t} V^T(\theta, \mathcal{M})\] is a difficult optimization problem

Used Gaussian process optimization with Monte Carlo estimate of expectation
Tuning the exploration schedule

Figure 2: Example estimated optimal decay schedules for $\epsilon$ in a Gaussian multi-armed bandit.
Controlling glucose levels

- A continuous-state MDP based on the mHealth study of Maahs et al. (2012) of patients with type 1 diabetes
  - Used previously by Luckett et al. (2018) for studying methods for estimating optimal infinite-horizon policy
- Covariates observed for patient $i$ at time $t$ is blood glucose level $Gl_i^t$, dietary intake $Di_i^t$, and physical activity $Ex_i^t$,
- Actions are give insulin or not, $A_i^t \in \{0, 1\}$
- Writing $x_i^t = (Gl_i^t, Di_i^t, Ex_i^t, A_i^t)$, glucose levels follow an AR(2) process:
  \[
  Gl^t = \beta_0 + \beta_1^T x^{t-1} + \beta_2^T x^{t-2} + e^t,
  \]
  \[
  e^t \sim N(0, \sigma^2)
  \]
- Food and activity generated iid at each time point
Controlling glucose levels

- Reward at time $t$ is a piecewise polynomial which increases as $GI^t$ departs from normal levels.
- For exploration strategy, we use $\epsilon$-greedy with a myopic policy.
  - i.e. The myopic policy is $A^t = \arg \max_a \hat{E}[U^t | x^t(a), x^{t-1}]$, where conditional mean function $\hat{E}[U^t | \cdot]$ is estimated with random forest.
Controlling glucose levels

Estimating generative model $\mathcal{M}$

- Three methods for estimating conditional distribution of glucose
  1. Correctly specified linear model (AR(2))
  2. Incorrectly specified linear model (AR(1))
     - $C^t \equiv$ estimated sampling distribution for linear model
  3. Correctly specified nonparametric conditional density estimator (NP)
     - $C^t \equiv$ bootstrap sampling distribution

- Distributions of food and activity estimated by their empirical distributions
Controlling glucose levels

<table>
<thead>
<tr>
<th>Time horizon</th>
<th>Method</th>
<th>Mean cumulative reward (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T=25</td>
<td>Tuned $\epsilon$-greedy (AR2)</td>
<td>-13.08(0.55)</td>
</tr>
<tr>
<td></td>
<td>Tuned $\epsilon$-greedy (AR1)</td>
<td>-13.39(0.84)</td>
</tr>
<tr>
<td></td>
<td>Tuned $\epsilon$-greedy (NP)</td>
<td>-30.40(1.97)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon$-greedy ($\epsilon = 0.05$)</td>
<td>-17.57(0.53)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_t = t^{-1}$</td>
<td>-20.02(1.25)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_t = 0.5t^{-1}$</td>
<td>-17.80(0.90)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_t = 0.8^t$</td>
<td>-22.93(1.79)</td>
</tr>
</tbody>
</table>

**Table 1:** Comparison of $\epsilon$-greedy variants in the glucose problem (n=15 subjects). 96 replicates for T=25, 192 replicates for T=50 (as these were considerably higher-variance).
## Controlling glucose levels

<table>
<thead>
<tr>
<th>Time horizon</th>
<th>Method</th>
<th>Mean cumulative reward (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T=50</td>
<td>Tuned $\epsilon$-greedy (AR2)</td>
<td>-16.92(0.42)</td>
</tr>
<tr>
<td></td>
<td>Tuned $\epsilon$-greedy (AR1)</td>
<td>-25.10(3.45)</td>
</tr>
<tr>
<td></td>
<td>Tuned $\epsilon$-greedy (NP)</td>
<td>-52.80(4.45)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon$-greedy ($\epsilon = 0.05$)</td>
<td>-39.56(2.55)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_t = t^{-1}$</td>
<td>-57.23(4.24)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_t = 0.5t^{-1}$</td>
<td>-33.83(3.41)</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_t = 0.8^t$</td>
<td>-52.93(4.45)</td>
</tr>
</tbody>
</table>

Table 2: Comparison of $\epsilon$-greedy variants in the glucose problem (n=15 subjects). 96 replicates for T=25, 192 replicates for T=50 (as these were considerably higher-variance).
Future work

Next step:
- Rules for deciding whether model estimate is good enough to guide exploration
- More thorough simulation study

Also would like:
- Modify for e.g. experimental settings where pure exploration is more important
- Theoretical regret analysis
- Alternative exploration methods which make full use of knowledge about the decision problem (incl time horizon and generative model)
Thanks!

Email: jclifto@ncsu.edu
