Outline

- Information collection and why it matters
- The challenge of learning
- An example: finding the best path
- Estimating beliefs
- The multi-armed bandit problem
Retrofitting buildings with new energy technologies

- Different combinations of technologies interact, with behaviors that depend on the characteristics of the building.
- Potential technologies include:
  - Window tinting, insulation
  - Energy-efficient lighting
  - Advanced thermostats
  - … many others
- We need to try different combinations of technologies to build up a knowledge base on different interactions, in different settings.
Information collection: finding the best path

- Figure out Manhattan:
  - Walking
  - Subway/walking
  - Taxi
  - Street bus
  - Driving
Information collection: finding effective compounds

Materials research

» How do we find the best material for converting sunlight to electricity?

» What is the best battery design for storing energy?

» We need a method to sort through potentially thousands of experiments.
Information collection: applications

- Pandemic disease control
  - Face masks are effective at disease containment.
  - It is better to test people for disease.
  - But we cannot test everyone. Who do we test?
Information collection: applications

- Finding good designs
  - How do we optimize the dimensions of tubes, plates and distances in an aerosol device?
  - Each design requires several hours to set up and execute.
  - Five parameters determine the effectiveness of the spray.

(c) 2013 W. B. Powell
Information collection: applications

- Tuning a medical experiment
  - Mice are injected with cancer cells, then with a new cancer drug
  - How many days should we wait to inject the drug in order to maximize its effectiveness?
Information collection: applications

- AFOSR contract to use optimal learning to advance the science of nano-biotechnologies
  » Joint with Peter Frazier at Cornell
  » Team of physical scientists at Princeton and Cornell

- Research challenges
  - Tuning concentrations, densities, temperatures, pressures, … using time-consuming physical experiments.
  - Anticipate learning well-defined belief models by exploiting understanding of underlying physics.
### Information collection: applications

- Experiments in materials science at Princeton
  - From research team of Prof. Sigurd Wagner, EE
    - 8-dimensional control vector (concentrations, temp, frequency)
    - Each observation requires a month of lab work

<table>
<thead>
<tr>
<th>Recipes</th>
<th>Stage I</th>
<th>Stage II</th>
<th>Two-stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta V_{T_1}$ (V)</td>
<td>$t_0$ (s)</td>
<td>$v$ (Hz)</td>
</tr>
<tr>
<td>C7 5/50/220, 2W</td>
<td>0.08</td>
<td>1.1</td>
<td>5e6</td>
</tr>
<tr>
<td>C5 5/50/220, 2W</td>
<td>0.2</td>
<td>0.4</td>
<td>5e6</td>
</tr>
<tr>
<td>T50 5/50/220, 2W</td>
<td>0.08</td>
<td>30.8</td>
<td>5e6</td>
</tr>
<tr>
<td>T13 6/60/300, 5W</td>
<td>0.08</td>
<td>10.3</td>
<td>5e5</td>
</tr>
<tr>
<td>T12 6/90/300, 5W</td>
<td>0.15</td>
<td>1.6</td>
<td>5e5</td>
</tr>
<tr>
<td>T14 6/120/300, 5W</td>
<td>0.14</td>
<td>1.3</td>
<td>5e5</td>
</tr>
<tr>
<td>Y3-2 10/150, 5W</td>
<td>0.1</td>
<td>&lt;&lt;1</td>
<td>5e4</td>
</tr>
<tr>
<td>Y3-3 6/160, 5W</td>
<td>0.1</td>
<td>&lt;&lt;1</td>
<td>1e5</td>
</tr>
<tr>
<td>Y3-1 10/100, 5W</td>
<td>0.3</td>
<td>&lt;&lt;1</td>
<td>5e3</td>
</tr>
</tbody>
</table>
Email from a research scientist (Nov. 2012)

» “A doctor has a time $T_{\text{orig}}$ to prescribe a drug (one of $N$ choices) to a patient in order to give the patient the best chance for survival. The doctor has some prior knowledge about which to prescribe.

» “The doctor can have a team of people run laboratory experiments. These experiments have costs (time, money,...) and some have more information for the problem at hand than others.

» “The doctor may be forced at any time $t < T_{\text{orig}}$ to make a decision.

» How should one approach this problem in a principled way? Would such a principled approach (of course also with approximation) be better than if we let the experts just use their experience to make a decision? Can we construct an algorithm that performs no worse than the experts?”
Information collection: areas of application

- Business processes
  - Pricing products
  - Revenue management
  - Finding the best vendor for supplying a component
  - R&D portfolio optimization
  - Identifying the ads that produce the most clicks
  - How do you find the retail stores with the highest rate of theft?
  - What are the best features for a laptop (or car)?

- Sports
  - What is the best starting lineup for a basketball team?
  - What are the four best rowers for a four-person shell?
  - Who are your best hitters? Who is best for the first four spots in the lineup?
Information collection: areas of applications

- **Health**
  
  » How do you find the best combination of drugs to treat high blood pressure or diabetes?
  
  » You would like to contain the spread of a disease such as H1N1 by having people wear face masks, but you cannot ask everyone to wear a face mask. By running tests, you can determine the population groups with the highest risk of the disease, but collecting this information takes time. How do you collect this information in the most effective way?
  
  » What is the best policy for testing people for heart disease or cancer?
  
  » You are trying to find the best molecular compound to fight cancer, but there are thousands of combinations. How do you sequence your tests in the most efficient way?
Information collection: areas of application

- **Energy**
  - You are trying to locate wind farms, but microgeography matters, which means you need to send inspectors to the field. How do you manage these inspectors as efficiently as possible?
  - You are trying to find the best material for photovoltaic cells (or batteries). Testing a compound takes time and money. How do you sequence your tests in the most effective way?
  - Tuning parameters for a device such as a hydrogen fuel cell or wind turbine – There are a number of physical parameters (height of the wind turbine, length of blades, weight, rate of rotation) that need to be tuned.
  - What is the best combination of energy saving technologies (insulation, tinted windows, motion sensors, computer-controlled thermostats) for a building.
  - What is the best portfolio of energy R&D technologies that the Department of Energy should be supporting?
Information collection: areas of application

■ The public sector

» A political candidate has to perform polling to determine where he/she should do campaigning. How should polling expenditures be allocated across states given a limited budget, to best support campaigning decisions?

» A university needs to know the likelihood of a certain type of student matriculating, but it has to actually make the offer to see whether a student will matriculate or not. Matriculation probabilities have a major impact on how a university should allocate its admission decisions.
Outline

- Information collection and why it matters
- The challenge of learning
- An example: finding the best path
- Estimating beliefs
- The multi-armed bandit problem
The challenge of learning

- Deterministic optimization
  - Find the choice with the highest reward (assumed known):

<table>
<thead>
<tr>
<th>Choice</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>759</td>
</tr>
<tr>
<td>2</td>
<td>722</td>
</tr>
<tr>
<td>3</td>
<td>698</td>
</tr>
<tr>
<td>4</td>
<td>653</td>
</tr>
<tr>
<td>5</td>
<td>616</td>
</tr>
</tbody>
</table>

The winner!
The challenge of learning

- Stochastic optimization
  
  » Now assume the reward you will earn is stochastic, drawn from a normal distribution. The reward is revealed after the choice is made.

<table>
<thead>
<tr>
<th>Choice</th>
<th>Mean</th>
<th>Std dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>759</td>
<td>120</td>
</tr>
<tr>
<td>2</td>
<td>722</td>
<td>142</td>
</tr>
<tr>
<td>3</td>
<td>698</td>
<td>133</td>
</tr>
<tr>
<td>4</td>
<td>653</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>616</td>
<td>102</td>
</tr>
</tbody>
</table>

The winner!
The challenge of learning

**Optimal learning**

» Now, you have a budget of 10 measurements to determine which of the 5 choices is best.

» You have an estimate of the performance of each, but you are unsure and you are willing to update your belief.

<table>
<thead>
<tr>
<th>Choice</th>
<th>Mean</th>
<th>Std dev</th>
<th>Observation</th>
<th>Mean</th>
<th>Std dev</th>
<th>Observation</th>
<th>Mean</th>
<th>Std dev</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>759</td>
<td>120</td>
<td>702</td>
<td>712</td>
<td>96</td>
<td>712</td>
<td>712</td>
<td>96</td>
<td>712</td>
</tr>
<tr>
<td>2</td>
<td>722</td>
<td>78</td>
<td>722</td>
<td>722</td>
<td>78</td>
<td>734</td>
<td>726</td>
<td>64</td>
<td>726</td>
</tr>
<tr>
<td>3</td>
<td>698</td>
<td>133</td>
<td>698</td>
<td>698</td>
<td>133</td>
<td>734</td>
<td>698</td>
<td>133</td>
<td>734</td>
</tr>
<tr>
<td>4</td>
<td>653</td>
<td>90</td>
<td>653</td>
<td>653</td>
<td>90</td>
<td>653</td>
<td>653</td>
<td>90</td>
<td>653</td>
</tr>
<tr>
<td>5</td>
<td>616</td>
<td>102</td>
<td>616</td>
<td>616</td>
<td>102</td>
<td>616</td>
<td>616</td>
<td>102</td>
<td>616</td>
</tr>
</tbody>
</table>

• … It is no longer obvious which you should try first.
The challenge of learning

At first, we believe that
\[ \mu_x \sim N(\mu_x^0, 1/\beta_x^0) \]

But we measure alternative \( x \) and observe
\[ W_X^1 \sim N(\mu_x, 1/\beta_\epsilon) \]

Our beliefs change:
\[ \mu_x \sim N(\mu_x^1, 1/\beta_\epsilon^1) \]
\[ \beta_x^1 = \beta_x^0 + \beta_\epsilon \]
\[ \mu_x^1 = \frac{\beta_x^0 \mu_x^0 + \beta_\epsilon W_X^1}{\beta_x^0 + \beta_\epsilon} \]

Thus, our beliefs about the rewards are gradually improved over \( n \) measurements

(c) 2013 W. B. Powell
The challenge of learning

- Now assume we have five choices, with uncertainty in our belief about how well each one will perform.
- If you can make one measurement, which would you measure?
The challenge of learning

- Now assume we have five choices, with uncertainty in our belief about how well each one will perform.
- If you can make one measurement, which would you measure?

(c) 2013 W. B. Powell
The challenge of learning

- Now assume we have five choices, with uncertainty in our belief about how well each one will perform.
- If you can make one measurement, which would you measure?

The value of learning is that it may change your decision.
The challenge of learning

- The measurement problem
  - We wish to design a *sequential* measurement policy, where each measurement depends on previous choices.

  - We can formulate this as a dynamic program:
    \[
    V^n(S^n) = \max_x \left( C(S^n, x) + E\left\{ V^{n+1}(S^{n+1}) \mid S^n \right\} \right)
    \]
    
    … but it is a little different than most dynamic programs that focus on the physical state.
The challenge of learning

- Optimal routing over a graph
  - \( S \) is a node in the network

\[
V^n(S^n) = \max_x \left( C(S^n, x) + E \left\{ V^{n+1}(S^{n+1}) | S^n \right\} \right)
\]
The challenge of learning

- Optimal routing over a graph
  » $S$ is a node in the network

$$V^n(S^n) = \max_x \left( C(S^n, x) + E\{V^{n+1}(S^{n+1}) \mid S^n\} \right)$$
The challenge of learning

- Learning problems
  - $S$ is our “state of knowledge”

$$S_5 = N(\mu_5, \sigma_5^2)$$

$$V^n(S^n) = \max_x \left( C(S^n, x) + E \left\{ V^{n+1}(S^{n+1}) \mid S^n \right\} \right)$$

- Hard to solve, because $S$ is now a multidimensional, continuous vector.
Dimensions of learning problems

Online vs. offline learning

» Offline learning – You have an opportunity to test different ideas before making your choice. You might be finding the best design in a laboratory, using simulation to find the best set of business processes, or running test markets before settling on a final design or price.
  • For example – the two trial runs for the OJ game would be an example of offline learning.

» Online learning – Learn as you go. Finding the best path to travel, finding the best medication for your high blood pressure, finding the best hitter, finding the best price for a product on the internet. You have to experience the choice to learn about it.
  • Once you start playing the OJ game, you have to learn as you go. This would be online learning.
Dimensions of learning problems

■ Objectives

» Maximizing reward/contribution/utility
  • Which prices/designs/choices perform the best?

» Minimizing expected opportunity cost
  • Captures the difference between how well you are doing (e.g. maximizing profits) and the best you could have done.

» Probability of correct selection
  • You want to maximize the probability you make the best choice.

» Indifference zones
  • We wish to maximize the likelihood that we pick a choice whose performance is in a zone that is acceptable.

» Minimizing fitting error
  • You might be trying to fit a statistical curve, and you want to minimize the error in your fit.
Outline

- Information collection and why it matters
- The challenge of learning
- An example: finding the best path
- Estimating beliefs
- The multi-armed bandit problem
Finding the best path

Finding the best path to work

» Four paths, but everyone time I drive on one, I sample a new time.

» I want to choose the path that is best on average.
Information acquisition

The shortest path game:

» Starting with the estimates at the top, choose paths so that you discover the best path.

<table>
<thead>
<tr>
<th>Path</th>
<th>Day</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Chosen path</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>25.00</td>
<td>24.00</td>
<td>22.00</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Information acquisition

- The shortest path game:
  » Starting with the estimates at the top, choose paths so that you discover the best path.

<table>
<thead>
<tr>
<th>Day</th>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
<th>Path 4</th>
<th>Chosen path</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.00</td>
<td>24.00</td>
<td>22.00</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Information acquisition

- The shortest path game:
  » Starting with the estimates at the top, choose paths so that you discover the best path.

<table>
<thead>
<tr>
<th>Day</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Chosen path</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.00</td>
<td>24.00</td>
<td>22.00</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Finding the best path

What do we know?

» The real average path times:
  » Mean time
    • Path 1 20 minutes
    • Path 2 22 minutes
    • Path 3 24 minutes
    • Path 4 26 minutes

  • Errors are +/- 10 minutes

» What we think:
  • Path 1 25 minutes
  • Path 2 24 minutes
  • Path 3 22 minutes
  • Path 4 20 minutes

» We act by choosing the path that we “think” is the best. The only way we learn anything new is by choosing a path.
Finding the best path

Illustration of calculations:

<table>
<thead>
<tr>
<th>Means</th>
<th>20</th>
<th>22</th>
<th>24</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spread</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stepsize</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Paths

<table>
<thead>
<tr>
<th>Day</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.91</td>
<td>17.29</td>
<td>14.43</td>
<td>33.46</td>
</tr>
<tr>
<td>2</td>
<td>20.49</td>
<td>12.90</td>
<td>30.19</td>
<td>16.43</td>
</tr>
<tr>
<td>3</td>
<td>20.77</td>
<td>22.57</td>
<td>32.90</td>
<td>17.61</td>
</tr>
<tr>
<td>4</td>
<td>22.50</td>
<td>18.69</td>
<td>25.67</td>
<td>16.72</td>
</tr>
<tr>
<td>5</td>
<td>19.08</td>
<td>16.38</td>
<td>19.15</td>
<td>25.18</td>
</tr>
</tbody>
</table>

Actual travel times

<table>
<thead>
<tr>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
<th>Path 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.00</td>
<td>24.00</td>
<td>22.00</td>
<td>20.00</td>
</tr>
<tr>
<td>25.00</td>
<td>24.00</td>
<td>22.00</td>
<td>26.73</td>
</tr>
<tr>
<td>25.00</td>
<td>24.00</td>
<td>24.73</td>
<td>26.73</td>
</tr>
<tr>
<td>25.00</td>
<td>23.64</td>
<td>24.73</td>
<td>26.73</td>
</tr>
<tr>
<td>25.00</td>
<td>22.65</td>
<td>24.73</td>
<td>26.73</td>
</tr>
</tbody>
</table>

Estimated travel times

Initial estimates of travel times

<table>
<thead>
<tr>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
<th>Path 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>24</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Finding the best path
Finding the best path

![Graph showing time vs iteration for different paths]
Finding the best path
Finding the best path
Finding the best path

■ What did we learn?
  » The initial estimate
  » How we react to new information
    • How quickly do we respond to new information?
    • How quickly should we respond?
  » Revisiting options we do not think are the best…

■ Some strategies
  » Go with the path you think is best
  » Pick a path at random
  » Go with the path you think is best, but occasionally strike out and try something else at random
  » ???
Outline

- Information collection and why it matters
- The challenge of learning
- An example: finding the best path
- Estimating beliefs
- The multi-armed bandit problem
Estimating beliefs

- Setup:
  Assume we are trying to estimate a mean $\mu$. Our current estimate (after $n$ observations) is
  \[ \theta^n = \text{Current estimate} \]
  And our latest measurement is
  \[ W^{n+1} = \text{Latest observation} \]

- Updating methods:
  - Frequentist – Uses only the observed data
  - Bayesian – Assumes that we have “prior knowledge” of the true values of the parameters.
Estimating beliefs

- Frequentist updating

A clumsy way to estimate means and variances using "frequentist" thinking is the standard formulas for means and variances:

\[ \theta^n = \frac{1}{n} \sum_{i=1}^{n} W^i \]  
A simple average

\[ \hat{\sigma}^{2,n} = \frac{1}{n-1} \sum_{i=1}^{n} (W^n - \bar{\theta}^n)^2 \]  
Remember to divide by \( n - 1 \)

\( \hat{\sigma}^{2,n} \) is an estimate of the variance of \( W^i \). Often we need an estimate of the variance of \( \bar{\theta}^n \). This is given by

\[ \bar{\sigma}^{2,n} = \frac{1}{n} \hat{\sigma}^{2,n} \]
Estimating beliefs

- **Frequentist updating**

  A more elegant way is to compute these statistics recursively. It is easy to verify that you get the same results using

  \[ \theta^n = \left(1 - \frac{1}{n}\right)\theta^{n-1} + \frac{1}{n}W^n \quad n = 1, 2, \ldots \]

  \[ \sigma^{2,n} = \begin{cases} 
  \frac{1}{n}(W^n - \theta^{n-1})^2 & n = 2 \\
  \left(\frac{n-1}{n-2}\right)\sigma^{2,n-1} + \frac{1}{n}(W^n - \theta^{n-1})^2 & n = 3, 4, \ldots 
  \end{cases} \]

  We still get the variance of \( \bar{\theta}^n \) using

  \[ \sigma^{2,n} = \frac{1}{n} \hat{\sigma}^{2,n} \]
Estimating beliefs

- Spreadsheet illustration

<table>
<thead>
<tr>
<th>n</th>
<th>thetahat</th>
<th>Direct</th>
<th>Recursive</th>
<th>Direct</th>
<th>Recursive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.713897</td>
<td>0.713897</td>
<td>0.713897</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.595058</td>
<td>0.654478</td>
<td>0.654478</td>
<td>0.007061</td>
<td>0.007061</td>
</tr>
<tr>
<td>3</td>
<td>0.234843</td>
<td>0.514599</td>
<td>0.514599</td>
<td>0.062229</td>
<td>0.062229</td>
</tr>
<tr>
<td>4</td>
<td>0.93743</td>
<td>0.620307</td>
<td>0.620307</td>
<td>0.086182</td>
<td>0.086182</td>
</tr>
<tr>
<td>5</td>
<td>0.004649</td>
<td>0.497175</td>
<td>0.497175</td>
<td>0.140444</td>
<td>0.140444</td>
</tr>
<tr>
<td>6</td>
<td>0.661151</td>
<td>0.524505</td>
<td>0.524505</td>
<td>0.116836</td>
<td>0.116836</td>
</tr>
<tr>
<td>7</td>
<td>0.952378</td>
<td>0.585629</td>
<td>0.585629</td>
<td>0.123517</td>
<td>0.123517</td>
</tr>
<tr>
<td>8</td>
<td>0.964887</td>
<td>0.633037</td>
<td>0.633037</td>
<td>0.123851</td>
<td>0.123851</td>
</tr>
</tbody>
</table>
Estimating beliefs

- **Bayesian updating**

Here we assume that we start with a prior mean and variance given by $\theta^0$ and $\bar{\sigma}^{2.0}$, where $\bar{\sigma}^{2.0}$ is our estimate of the variance of $\theta^0$. Now assume that when we make a measurement, the noise in the measurement is known to have mean 0 and variance $(\sigma_w)^2$. To write the updating equations, it is convenient to define

$$\beta^n = (\bar{\sigma}^{2,n})^{-1} \quad \beta_\varepsilon = (\sigma_w)^{-1}$$

Using this assumed information, the updating formulas become

$$\beta^n = \beta^{n-1} + \beta_w$$

$$\theta^n = \frac{\beta^{n-1} \theta^{n-1} + \beta_w W^n}{\beta^{n-1} + \beta_w}$$
Outline

- Information collection and why it matters
- The challenge of learning
- An example: finding the best path
- Estimating beliefs
- The multi-armed bandit problem
Multi-armed bandits

- $M$ slot machines
  - We do not know the mean winnings from each slot machine
  - We can collect information by playing a machine
  - Did we win because this machine has a higher winning probability, or did we just get lucky?
  - Another machine with lower performance might actually be better. But how do we balance our efforts?
Multi-armed bandits

What decision do we make?

» *Exploitation:* play the one we think is best
  
  - We have reason to believe that this alternative will give us a good reward

» *Exploration:* just try something new and learn about it
  
  - Sacrifice some immediate gain, but make better decisions in the future
Multi-armed bandits

- Remember the difference between offline and online problems...

**Offline:**
Collect information in a lab setting or a simulator, then make a decision

**Online:**
Learn in real-time, pay a cost for making a poor choice

- The bandit problem is online – if we play the wrong slot machine, we lose money
- What about the OJ Game...?
Measurement policies

Measurement policies for discrete problems

» Pure exploration – Make choices at random so that you are always learning more.

» Pure exploitation – Always make the choice that appears to be the best.

» Hybrids
  • Explore with probability $\epsilon$ and exploit with probability $1 - \epsilon$
  • Epsilon-greedy exploration – explore with probability $\epsilon^n = c / n$. Goes to zero as $n \rightarrow \infty$, but not too quickly.
Measurement policies for discrete problems

- **Boltzmann exploration**
  
  - Explore choice $x$ with probability
    
    $$p_x^n = \frac{\exp(\alpha \theta_x^n)}{\sum_{x'} \exp(\alpha \theta_{x'}^n)}$$
  
  - Sometimes called a “soft max” policy.

- **Interval estimation**
  
  - Choose $x$ which maximizes $\mu_x^n + z_\alpha \overline{\sigma}_x^n$

- **Upper confidence bounding**
  
  - Choose $x$ which maximizes
    
    $$\nu_x^{U CB-normal,n} = \theta_x^n + 4\sigma_w \sqrt{\frac{\log n}{N_x^n}}$$

  where $N_x^n$ = number of times $x$ has been tested
Interval estimation

» Choose the option with the highest value of

\[ \nu^{IE,n}_x = \theta^n_x + z_\alpha \bar{\sigma}^n_x \]

Uncertainty bonus

where \( z_\alpha \) is the number of standard deviations in a normal distribution leaving a probability \( \alpha \) in the tail. \( \bar{\sigma}^n_x \) is the standard deviation of \( \theta^n_x \).

The value \( z_\alpha \) is treated as a tunable parameter, and it is important that it is tuned carefully. Common practice is to choose values around 2 or 3 (so we are evaluating an alternative based on 2 or 3 standard deviations above the mean).
Multi-armed bandits

- Gittins indices
  » In 1974, Gittins and Jones showed that all you had to do was to try the choice with the highest index given by

\[ \nu_{x}^{\text{Gitt},n} = \theta_{x}^{n} + \Gamma_{x}(n)\sigma_{W} \]

where \( \bar{\theta}^{n} \) is the current estimate of the mean, and \( \sigma_{W} \) is the standard deviation of the noise (not the std. dev of \( \theta_{x}^{n} \)).

\( \Gamma(n) \) is the “Gittins index” if the mean is 0 and variance is 1.

» What is \( \Gamma(n) \)? That is the Gittins magic. This needs to be computed using a fairly difficult calculation.

» This is known as an index policy, because choosing the best option means choosing the one with the highest index.

» This works only for a very special case – discounted, infinite horizon, independent observations, stationary, ….
Multi-armed bandits

- Gittins indices
  - The table to the left gives the index $\Gamma(n)$ for mean 0, variance 1.
  - Note that the index declines with the number of observations, approximately with $1/\sqrt{n}$
  - But, computing these values is hard, and using tables is clumsy (and restrictive)

<table>
<thead>
<tr>
<th>Observations</th>
<th>Discount factor Known variance</th>
<th>Discount factor Unknown variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>1</td>
<td>0.9956</td>
<td>1.5758</td>
</tr>
<tr>
<td>2</td>
<td>0.6343</td>
<td>1.0415</td>
</tr>
<tr>
<td>3</td>
<td>0.4781</td>
<td>0.8061</td>
</tr>
<tr>
<td>4</td>
<td>0.3878</td>
<td>0.6677</td>
</tr>
<tr>
<td>5</td>
<td>0.3281</td>
<td>0.5747</td>
</tr>
<tr>
<td>6</td>
<td>0.2853</td>
<td>0.5072</td>
</tr>
<tr>
<td>7</td>
<td>0.2528</td>
<td>0.4554</td>
</tr>
<tr>
<td>8</td>
<td>0.2274</td>
<td>0.4144</td>
</tr>
<tr>
<td>9</td>
<td>0.2069</td>
<td>0.3808</td>
</tr>
<tr>
<td>10</td>
<td>0.1899</td>
<td>0.3528</td>
</tr>
<tr>
<td>20</td>
<td>0.1058</td>
<td>0.2094</td>
</tr>
<tr>
<td>30</td>
<td>0.0739</td>
<td>0.1520</td>
</tr>
<tr>
<td>40</td>
<td>0.0570</td>
<td>0.1202</td>
</tr>
<tr>
<td>50</td>
<td>0.0464</td>
<td>0.0998</td>
</tr>
<tr>
<td>60</td>
<td>0.0392</td>
<td>0.0855</td>
</tr>
<tr>
<td>70</td>
<td>0.0339</td>
<td>0.0749</td>
</tr>
<tr>
<td>80</td>
<td>0.0299</td>
<td>0.0667</td>
</tr>
<tr>
<td>90</td>
<td>0.0267</td>
<td>0.0602</td>
</tr>
<tr>
<td>100</td>
<td>0.0242</td>
<td>0.0549</td>
</tr>
</tbody>
</table>
Multi-armed bandits

- An approximation for Gittins indices
  » From Chick and Gans (2009)
    - First compute:
      \[
      \tilde{b}(s) = \begin{cases} 
      \frac{s}{\sqrt{2}} & s \leq \frac{1}{7} \\
      e^{-0.02645 \log^2 s + 0.89106 \log s - 0.4873} & \frac{1}{7} < s \leq 100 \\
      \sqrt{s} \left( 2 \log s - \log \log s - \log 16\pi \right)^{\frac{1}{2}} & s > 100.
      \end{cases}
      \]

    - The Gittins policy is then given by
      \[
      X^{Gitt,n}(S^n) = \arg \max_x \left\{ \theta^n_x + \sigma_w \sqrt{-\log \gamma \tilde{b}} \left( -\left( \frac{\sigma^n_x}{\sigma^n_x} \right)^2 \right) \right\}
      \]

      - \(\mu^n_x = \) Estimate of the value of \(x\) after \(n\) measurements
      - \(\left( \sigma^n_x \right)^2 = \) Estimate of variance of \(\theta^n_x\) after \(n\) measurements
      - \(\left( \sigma^n_w \right)^2 = \) Variance of a measurement (assumed known)
      - \(\gamma = \) Discount factor
Multi-armed bandits

- Upper confidence bounding
  - Choose the arm based on a probabilistic upper bound
  - If we have normal rewards, this would be
    \[ \nu_{x}^{\text{UCB-normal}, n} = \theta_{x}^{n} + 4\sigma_{w} \sqrt{\frac{\log n}{N_{x}^{n}}} \]
    where \( N_{x}^{n} \) = number of times \( x \) has been tested
  - The policy is to choose \( x \) with the largest \( \nu_{x}^{\text{UCB-normal}, n} \)
  - UCB policies tend to be more conservative, but are more robust with respect to violations of distributional assumptions.
  - UCB policies tend to be used in applications where measurement budgets are large.