Optimal Decision Making in the Presence of Uncertainty: Managing Disease Outbreaks

Sandya Lakkur
sandya.s.lakkur@vanderbilt.edu

Department of Biostatistics
Vanderbilt University

August 8, 2016

Contents

1 Introduction: Foot and Mouth Disease

Foot-and-mouth disease is a rapidly spreading virus that affects cloven-hoofed animals, not to be mistaken with hand, foot, and mouth disease which affects humans. It is transmitted through direct contact with an infected animal or aerosols. This disease has an average incubation period of four days and can survive in the animal's environment for over a month (depending on pH and temperature). While this virus is not deadly, it can cause: blisters in the mouth and feet, drop in milk production, weight loss, and lameness.[3] These symptoms take a toll on a farm's financial gain, thus infected animals are usually culled. This disease containment strategy had drastic consequences during the 2001 outbreak in the United Kingdom. By the end of the outbreak, over six million livestock were culled which resulted in a 3.5 billion dollar cost to the UK agriculture and food industry. This excludes the 2.8 billion dollars paid by the government to handle disposal and cleanup costs.[4] Due to mismanagement of the disease spread the epidemic was not quickly eradicated. Some specific examples of mismanagement included: not having an adequate number of veternarians to treat infected livestock, implementing a national movement ban too late into the outbreak, and misunderstanding what information should be collected in the field regarding disease spread.[10] Because there were large economic and environmental tolls from this outbreak, there became a need to explore how to prevent a future epidemic of this magnitude. Ultimately, timely and informed decisions should have been made during the disease outbreak, and clear communication between field experts and agricultural managers needed to occur. This event inspires the need to apply optimal decision making. In this discussion, first the framework of decision analysis will be explained. Then an extension to decision analysis, adaptive management, will be explored. Once this foundation has been established, the method of finding an optimal policy, reinforcement learning, will be introduced. The discussion will then conclude with exploring optimal decision making in the 2001 foot-and-mouth disease outbreak, and ideally provide an understanding of how to prevent another management disaster.

2 Decision Analysis

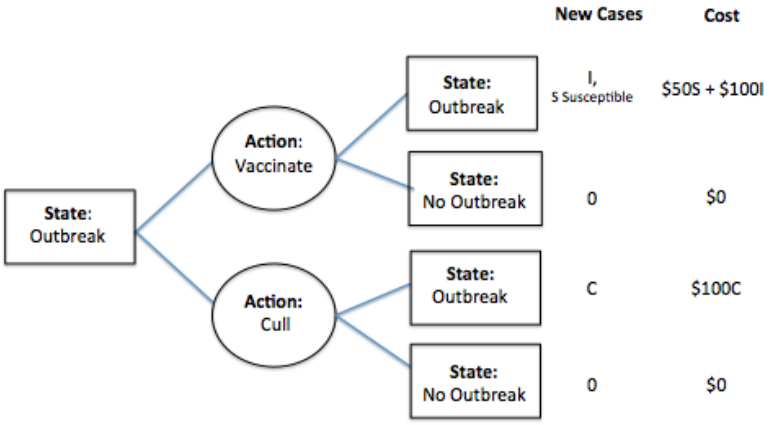
2.1 The Decision Making Framework

Decision analysis can be formally defined as: a quantitative method of evaluating management options using information about uncertainties.[8] This definition can be deconstructed into two main components: evaluating management options and uncertainty. Evaluating management options requires knowledge about the following: the objective, actions, states, and rewards. Before any management strategies can be evaluated an objective (or objectives) needs to be defined. The objective provides a metric to judge different decisions. In the context of foot-and-mouth disease, suppose the objective is to minimize the duration of the outbreak. Competing management strategies can resultantly be evaluated based on how long the outbreak persists under each respective decision. The objective(s) also defines the environment - which consists of states, actions, and rewards- for the decision making process. States refer to the measurable quantities or qualities of the environment in which a decision is made. Naturally the more states an environment has, the more complicated the decision. Actions refer to the possible management options a decision maker can execute. As with states, the more actions a decision maker can choose from the more complicated the problem. Each state (or each state-action pair) is associated with a specific reward. This reward provides a summary of how desirable it is to be in a specific state.[9] The higher the reward, the more desirable the state. These quantities are usually defined a priori through previous knowledge. Rewards may also be used interchangeably with costs if the objective involves some expense from the decision maker. One can now discuss objectives in a more generic way: the objective, in a decision-making problem, is to determine which decision maximizes the reward, or minimizes the cost. To better understand this framework two examples in the context of the foot-and-mouth disease outbreak are presented.

For the first example, suppose the objective is to minimize the cost of livestock lost during an outbreak to farmers. Costs will only be incurred if there is an outbreak. This objective naturally establishes the states for which a decision must be made: there is currently an outbreak, and there is currently no outbreak. Notice that in this example there are only two states which implies that this decision is not particularly complex. Suppose that in this example, the initial state of the

environment is: there is currently an outbreak. To determine whether or not the outbreak continues at the next time point, also referred to as determining the next state, an action needs to be implemented. It is important to note that an action need not be an intervention; doing nothing is also an action that causes an initial state to evolve into another state. Since foot-and-mouth disease travels quickly between farms, due to movement of livestock, a common management decision is ring culling: killing any livestock within a certain radius of an infected farm. Suppose that this strategy is replicated with vaccinations: vaccinate all susceptible livestock within the pre-determined ring and cull only the infected. This segues to the possible management actions: ring culling, and ring vaccination with only culling the infected. In this particular example costs will be used instead of rewards, because the objective is to minimize cost to farmers. Recall the contribution of costs to the decision-making problem: the costs provide a summary of how undesirable it is to be in a particular state (or state-action pair). That is, the higher the cost the less desirable it is to be in a state - or follow a particular action state sequence. Suppose that it costs \$100 to cull one livestock and \$50 to vaccinate one livestock. This means that a farmer would have to spend \$100C under the culling strategy, where C is the number of livestock to be culled within the radius of the infected farm. Under the second strategy a farmer would need to spend \$50S + \$100I, where S represents the number of susceptible livestock in the ring and I represents the number of infected livestock. The following decision tree represents the state, action, state sequence with associated costs:

Figure 1: Example 1 Decision Tree



An outbreak in this example is loosely defined as having a nonzero number of infected cases. Thus, any terminal node in the decision tree with a “no outbreak" state will have zero new cases. The costs are calculated by multiplying the number of infected -or susceptible- cases with the associated culling and vaccination costs, respectively. Since the objective is to minimize costs to the farmer, the optimal action is chosen by calculating the average cost across the different states. Notice the role of costs in this example: under both actions, it is less desirable to be in the subsequent “outbreak" state since its associated cost is higher than that of the subsequent “no outbreak" state. This motivates the need to choose an action that is more likely to result in a subsequent “no outbreak" state.

The decision process can be much more complex when there are many potential states in the environment. This leads to the second example; consider elements from the first example with some slight changes:

- Objective: Minimize cost of livestock lost to farmers
- Actions: (1) Vaccinate susceptibles and cull infected in a pre-determined ring, (2) ring culling
- States: Number of infected livestock, number of susceptible livestock
- Cost: \$100 per livestock cull, \$50 per livestock to vaccinate

Notice that in this example there are two discrete states, whereas in the previous example there was one binary state. Constructing the decision tree for this example would be impractical. If there were N livestock in this example then each action would result in N+1 different combinations of states, resulting in 2(N+1) terminal nodes since there are two management actions. The complexity can increase again if there more actions to choose from; if there are N actions then the decision tree would include N(N+1) terminal nodes. Visualizing a tree this large is difficult, let alone extracting any valuable information from it. For these types of environments, it is more convenient to think of evaluating decisions by optimizing a function rather than optimizing a decision tree.

2.2 Average Utility

In the two previous examples decision trees were introduced to help visualize the complexity of the decision problem. The next step is to actually make a decision, and to do this the average cost of a management action needs to be calculated. Average cost is defined as:

$$\bar{R}(a) = \sum_{s} P(s \mid a) R(a, s) \tag{1}$$

To prevent ambiguity between costs and rewards, $\bar{R}(a)$ will the denote average utility of an action for the remainder of this discussion. Defining some notation: a represents action, s represents the subsequent state, and R represents the utility. $P(s \mid a)$ is known as a transition probability, the probability that the subsequent state s is observed given that action a is implemented. These transition probabilities are defined a priori through previous studies or other expert opinions. Deconstructing the components of (1) average utility has a very intuitive interpretation: the average utility of an action a is the sum of the utilities of the resulting states weighted by the states' transition probabilities.

Applying this definition to the first example, the decision tree can be edited to include the transition probabilities. Just for illustration, the transition probabilities are chosen randomly.

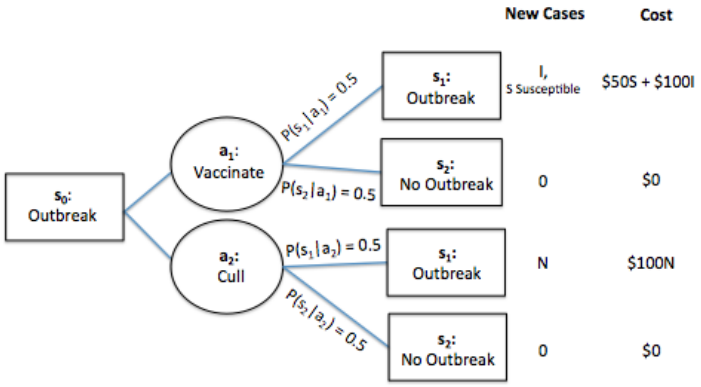


Figure 2: Example 1 Complete Decision Tree

Notice that the transition probabilities displayed in the above tree suggest the outbreak and no outbreak states are equiprobable under each management action. This means that the decision cannot be made by visual insepction, i.e. - determining which action has a higher probability of resulting in the 'no outbreak" state. Thus equation (1) can be applied to compare average costs of the two actions: $\bar{R}(a_1) = 25S + 50I$ and $\bar{R}(a_2) = 50N$. Based on these quantities, to meet the objective of minimizing the cost of livestock lost during an outbreak to the farmer, the optimal action would be based on which of the two average utilities (25S + 50I, 50N) is smaller. The same process can be implemented for the second example; instead of taking an average across two states an average would need to be calculated over N+1 states. In this particular example, it was assumed that the transition probabilities represented the truth. Although transition probabilities are based on expert opinion and previous knowledge, there is still uncertainty associated with these estimates.

2.3 Types of Uncertainty

Recall that decision analysis incorporates both a mechanism for evaluating management strategies, through average utility, while incorporating information about uncertainties. Two forms of uncertainty arise in management problems: aleatoric and epistemic. Aleatoric uncertainty refers to variability or stochastic uncertainty that cannot be eliminated.[11] Epistemic uncertainty refers to the uncertainty generated due to a lack of knowledge of the processes in the environment. One way this can be reduced is through investment of monitoring the state action sequence.[6] There are two main forms of aleatoric uncertainty: stochasticity (environmental variation) and partial controllability. Stochasticity refers to the uncontrollable factors theat may cause an action to result in an unplanned state. Partial controllability refers to the the degree to which the proposed action is correctly implemented in the population. While these two forms of aleatoric uncertainty cannot be eliminated by definition they can, and should, be accounted for. Estimating aleatoric uncertainty is essential for understanding the variability of an action. That is, an action can be simulated many times operating under the same conditions, and the resulting state may not be the same every time. Without accounting for this uncertainty the decision maker must assume that an action will always result in the same subsequent state. In example 1 the transition probabilities represented a function of stochasticity and partial controllability. The probability that an outbreak will occur given that the vaccination of susceptible livestock and culling of infected livestock strategy is implemented is 50%. Notice that if this quantity was exactly zero, the decision-maker would be assuming that the vaccination strategy perfectly prevents an outbreak. However, the nonzero quantity suggests that factors such as climate, livestock movement, and ineffective vaccinations could be accounted for. Transition probabilities could better represent the truth based on investment in monitoring (e.g. the probability of a vaccination strategy resulting in an outbreak may actually be 40% instead of 60%), however it will not be zero or one unless the management strategy is completely effective or completely ineffective

The two main forms of epistemic uncertainty are: partial observability and process uncertainty. Partial observability refers to the sampling variation in the monitoring of states. Process uncertainty refers to the uncertainty in a manager's assumptions of a decision- outcome relationship. Recall in example 1 that environmental variation and partial controllability were accounted for through the transition probability. Based on expert opinion and previous knowledge (note that the transition probabilities in this example were chosen randomly, ideally expert opinion and previous knoweldge would be used to generate an estimate of aleatoric uncertainty), outbreak and no outbreak states were equiprobable under the vaccination strategy (i.e: $P(Outbreak \mid Vaccination) = P(No\,outbreak \mid Vaccination) = 0.5$). However, it is possible to incorporate uncertainty about this assumption through specification of a competing assumption. Suppose that the competing assumption in example 1 is: a vaccination scheme results in a higher probability of no outbreak, or in other words $P(No\,outbreak \mid Vaccination) > P(Outbreak \mid Vaccination)$. Operating under the second assumption creates a slightly different decision tree; the states, actions, and utility remain the same however the transition probabilities would be different. An optimal decision in this context could be determined by drawing out both decision trees and using equation (1) to determine which action was best. Instead this process can be expressed formulaically through the following modification to equation (1):

$$E[\bar{R}(a)] = \sum_i p(H_i)\bar{R}_i(a) \tag{2}$$

This equation also has a straightforward interpretation: the expected utility is simply the average utility for an action weighted by the belief in a hypothesis, or assumption. Also notice the added layer of complexity by accounting for this uncertainty. If there are many hypotheses to choose from, many states, and many actions this problem becomes more complicated. This phenomenon is referred to as the curse of dimensionality. It is formally defined as: the number of operations needed to arrive at an optimal decision tends to grow exponentially with the number of state variables.[9] This motivates the need for more efficient optimization algorithms, which will be discussed in the reinforcement learning section. To define some notation for equation (2), $p(H_i)$ represents the probability of belief in a specific hypothesis, H_i. H_1 could represent the hypothesis that an outbreak and no outbreak are equiprobable under the vaccination strategy, and H_2 could represent the hypothesis that the vaccination strategy results in a higher probability of no outbreak. The $p(H_i)$ quantities are also

decided by the decision maker a priori, however they are chosen based on best guesses. This can affect the decision-making process because, if there is a strong belief in the first hypothesis when the second hypothesis represents the truth then the average utilities will not be correct. As mentioned previously, investment in monitoring is one method that can improve these estimates. Through use of adaptive management, the information accumulated from monitoring can aid in improved decision-making. This will be further discussed in the next section.

In reference to example 2 it is likely that the true number of infected and susceptible livestock were not observed, and decisions were made based on imperfect observations. This is an example of partial observability. The estimate of the subsequent state based on an action becomes affected because the original state was not correctly represented. A previous study found that, in the case with discrete states, partial observability could be investigated through a sensitivity analysis (Yaesoubi, Cohen et al, 2009). The estimates for the subsequent number of infected can be quantified as: $\hat{I}_{t+1} = (1+e(t))I_t$, with a similar form for the number of subsequent susceptibles. The term e(t) represents the error associated with identifying the correct number of cases at time t, and I_t represents the number of cases identified at time t. Ideally the term e(t) would be known to the decision maker, and the estimate of the number of infected at time t could be immediately adjusted. Since this is normally not the case, a distribution can be imposed on e(t) and various values of error can be generated. A sensitivity analysis could be performed to examine how decisions change based on different random errors.

3 Adaptive Management

3.1 Bayesian Framework

So far this discussion has deconstructed decision analysis into average utility and different forms of uncertainty. Equation (2) was introduced as a means of arriving at an optimal decision. These pieces are all necessary to understand how and why adaptive management works. Adaptive management can be formally defined as: a systematic process for continually improving management policies and practices by learning from the outcomes of operational programs.[6] This process can essentially be thought of as a sequential analysis- a method where data is evaluated as it is being collected. However, the distinguishing factor in adaptive management is that beliefs are updated as data is collected. Because a main component of adaptive management is learning based on accumulated information, a Bayesian framework is best suited for this problem. Before delving into the application of the Bayesian framework to adaptive management, let us first revisit the Bayesian paradigm.

In the Bayesian paradigm, conclusions about a parameter of interest are made in terms of probability statements. This differs from the frequentist paradigm where conclusions are based on the procedure used to estimate the parameter of interest.[5] Essentially, prior knowledge of the parameter of interest is incorporated with information obtained from the data by means of Bayes rule. Suppose that θ is the parameter of interest and y represents the data collected from an experiment. Bayes rule states that:

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)} \propto p(\theta)p(y \mid \theta) \tag{3}$$

The quantity $p(\theta \mid y)$ refers to the posterior distribution; it represents the distribution of θ conditional on the information collected, y. Using this derived distribution, inferences can be made on the parameter of interest. The terms $p(\theta)$ and $p(y \mid \theta)$ refer to the prior density and data likelihood respectively. It is important to notice that the posterior distribution is simply the prior weighted by the likelihood. This weight can either be informative and overwhelm the information collected from the experiment, or non-informative allowing for the data to explain most of the variation. Sensitivity analyses can usually be performed to assess the choice of prior distribution. Once the posterior distribution can be obtained, which may not always be a simple task, parameter variance and expectation can be directly computed. This can lead to the construction of a credible interval which is interpreted using probability statements.

3.2 Application to Adaptive Management

A central goal in adaptive management is learning about the environment through accumulated data; data in the decision analysis context refers to the state of the environment. This goal lends itself well in reducing epistemic uncertainty. Recall that a decision maker's probabilities of belief in hypotheses could severely affect the expected utility of an action, resulting in a potentially suboptimal and misinformed decision. Using a more formal explanation, adaptive management can reduce epistemic uncertainty by making the following changes to equation (2):

$E[\bar{R}(a)] = \sum_i p(H_i)\bar{R}_i(a)$

$= \sum_i p(H_i)[\sum_s p_i(s \mid a)R_i(a,s)]$

Aside: $p'(H_i) = \frac{p(H_i)p_i(s|a)}{\bar{p}(s|a)}$, where $\bar{p}(s \mid a) = \sum_i p(H_i)p_i(s \mid a)$

$= \sum_i \frac{p'(H_i)\bar{p}}{p_i(s|a)}[\sum_s p_i(s \mid a)R_i(a,s)]$

$= \sum_s \bar{p}(s \mid a)\bar{R}[a, s \mid p'(H_i)]$, where $\bar{R}[a, s \mid p'(H_i)] = \sum_i p'(H_i)R_i(a,s)$

In the third line, the term $p(H_i)$ is re-weighted based on the degree to which the hypothesis H_i correctly predicted the future state. Notice that $p'(H_i)$ is simply an application of Bayes rule where the prior is represented by the current belief in hypothesis H_i. This process is referred to as likelihood updating [12]. As more data is collected, $p'(H_i)$ will continually update, ideally resulting in convergence to the truth. The steps of implementing adaptive management can be summarized as:

1. Initialize the beliefs for each hypothesis
2. Determine the current state of the system via monitoring
3. Apply optimal action based on results from average utility function
4. Monitor the system and collect state information
5. Learn about $p(H_i)$ using $p'(H_i) = \frac{p(H_i)p_i(s|a)}{\bar{p}(s|a)}$
6. Apply the optimal action based on results from the updated average utility function
7. Repeat steps 4-6 as necessary

The method outlined above is known as active adaptive management. Essentially the best management option is chosen while accounting for uncertainty in assumptions, and learning is a key component. This method is particularly attractive because decision-makers have limited resources. Addressing uncertainty presents an opportunity for the decision maker to adjust how resources are allocated at different time points. It is possible for a static management action, one that does not change throughout the decision making process, to overallocate resources to achieve the objective. One challenge with active adaptive management is balancing the degree of learning to improve future management, with achieving the best short-term outcome based on current knowledge. The value of learning can be evaluated using the expected value of perfect information (EVPI). EVPI estimates the utility to the decision maker of addressing one or more uncertainties prior to the implementation of specific decisions. It is calculated by computing the average of optimal expected utilities (optimal expected utility assumes that the sequence of actions that optimizes utility is followed), assuming that each of the i hypotheses are correct, compared to the utility averaged over all hypotheses.[2] This allows for an intuitive interpretation: if EVPI is equal to 0 then there is no value in learning which hypothesis about the environment is correct.

3.3 Case Studies of Adaptive Management

Adaptive management has an intuitive benefit to decision-makers: accumulating data results in better decision making. The following discusses two studies where the incorporation of adaptive management yielded more informative decision-making. The goal of the first study was to explore the differences in the trajectory of an influenza outbreak and the expected cost of implementing an intervention under a static management strategy and an adaptive management strategy.(Merl, Johnson, et al, 2009) The actions available to the decision makers included: percentage of the susceptible population to vaccinate (α) and the threshold at which the number of susceptibles must be at to stop the vaccination campaign (γ). At each time step in the outbreak Markov Chain Monte Carlo (MCMC) was used to estimate transmission rate, recovery rate, and mortality rate. Notice that through the construction of these distributions, the decision maker has a means to account for process uncertainty. More specifically, conditional on each of the likely parameter values, the epidemic was simulated forward 100 times via binomial processes and Monte Carlo methods were implemented to determine the combination of α and γ that minimized the expected cost. It was concluded that the expected costs under an adaptive intervention were smaller in comparision to the static intervention, while the trajectory of the epidemic was similar under both interventions. It is important to note that the approach taken here, of specifying a statistical model to estimate uncertainty and optimizing through Monte carlo methods, is not the only approach that could have been used to make optimal decisions.

Another study assessed the effect of adaptive policies during an influenza epidemic by comparing the disease trajectories controlling for different levels of a decision-maker's willingness to pay for an intervention (Yaesoubi, Cohen et al, 2011). Influenza epidemics were simulated 1000 times through Poisson processes assuming three different levels of willingness to pay. When a decision-maker implemented a more rigorous intervention, had a higher willingness to pay, the expected number of infected individuals was smaller compared that of a less rigorous intervention. It was concluded that an adaptive intervention would be an appealing approach because the intervention rigor could be adjusted based on how far along the disease outbreak has progressed. This investigation also highlights that adaptive management addresses decision making under resource constraints. The example used to demonstrate this described a case where there were n vaccines available at the beginning of the epidemic, and the decision maker needed to decide how many susceptibles to vaccinate using the n vaccines. It was assumed that no additional vaccines would become available during the epidemic. The objective in this example was to minimize cost, and uncertainty was incorporated through different levels of a decision-maker's willingness to pay for an intervention. A dynamic programming approach was used here: the optimal number of vaccines was determined through a backward iterative algorithm, conditional on the initial proportion of susceptibles. The vaccination strategy was assessed by its associated monetary cost. While both of these studies highlighted that adaptive interventions could reduce monetary costs to the decision-maker it would be interesting for the EVPI to included. This way the effect of learning, adjusting for uncertainty, could be assessed.

4 Optimizing the Expected Utility

In the previous section, two case studies illustrated the monetary benefits to decision-makers in implementing adaptive interventions. Prior to that, adaptive management was developed through defining the expected utility function. The next piece is to define how to optimize the utility function in more complex settings. The examples presented previously have been simple: Example 1 had one binary state, and Example 2 had two discrete states but was concerned with making a decision at only a single timepoint. Since adaptive management has been introduced, optimal decision making can now occur at multiple timepoints while also addressing uncertainty. Let i represent the different levels of process uncertainty, or different hypotheses about the environment. Then the i^{th} expected utility function can be defined as:

$$Q_i(s_t, a_t) = E[\sum_{j=t}^{T} R(a_j \mid s_j) \mid s_t] \tag{4}$$

Notice that the expected utility is defined using a Q, this is known as the action-value function under belief i. For the remainder of this discussion Q will be used to refer to the expected utility calculated at one or more time points. Also

notice that equation (4) is essentially equation (1) with an additional summation taken with respect to time. In equation (1) $\bar{R}$ was used to represent the average utility weighted by transition probabilties. In equation (4) the average utility is again weighted by transition probabilties but for the remainder of the discussion an expectation will be used to represent an average utility. Optimal decision making considers the long term reward which implies that an expectation is more fitting than a bar. Equation (4) can be written recursively through the following re-expression:

$$Q_i(s_t, a_t) = E[R(a_t \mid s_t) + \textstyle\sum_{j=t+1}^{T} R(a_j \mid s_j) \mid s_t]$$

$$= R(a_t \mid s_t) + \textstyle\sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) E[\sum_{j=t+1}^{T} R(a_j \mid s_j) \mid s_{t+1}]$$

$$= R(a_t \mid s_t) + \textstyle\sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) Q_i(s_{t+1}, a_{t+1})$$

This simplification has an intuitive interpretation: the first term represents the immediate utility of an action (at time t), and the second term represents the long term expected utility of an action. The recursive form of equation (4) is referred to as the Bellman equation. Instead of expressing a different Bellman equation for each i^{th} hypothesis, the hypothesis weights can be incorporated into the Bellman equation directly:

$$\bar{Q}(s_t, a_t) = \sum_i P(H_i)[R(a_t \mid s_t) + \sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) Q_i(s_{t+1}, a_{t+1})] \tag{5}$$

In examples 1 and 2 optimization was carried out directly by comparing the expected utilities for each action, but was never formally defined. The following addition can be made to equation (5) to introduce optimization:

$$\pi(s) = argmax_{a_t}[\sum_i P(H_i)[R(a_t \mid s_t) + \sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) Q_i(s_{t+1}, a_{t+1})]] \tag{6}$$

Equation (6) refers to a policy (expressed using π); it provides a mapping from a current state to an action. This function is the deliverable to the decision-maker. In a simple decision problem the policy could be a lookup table, however in a more complex decision problem the policy may require more extensive computation. This discussion, so far, has suggested that the goal of decision analysis is to generate an optimal decision under uncertainty. While this is true, the goal can be more formally expressed as: generating the optimal *policy*, a function, while incorporating uncertainty.

An optimal policy π^* is considered better than or equal to another policy π if its expected utility is greater than or equal to that of π for all states in the environment.[9] The optimal action-value function $\bar{Q}^*$, also referred to as the Bellman optimality equation, calculates the expected return for taking action a from state s, and assumes that optimal actions are subsequently implemented. It is expressed as:

$$\bar{Q}^*(s_t, a_t) = max_{a_t}[\sum_i P(H_i)[R(a_t \mid s_t) + \sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) Q_i(s_{t+1}, a_{t+1})]] \tag{7}$$

Once the Bellman optimality equation is specified, it is trivial to determine the optimal policy. For each state in the environment there will be at least one action at which the Bellman optimality equation is maximized. This generates the optimal policy. It is important to note the the Bellman optimality equation is a system of s equations, one for each state. In order to generate the optimal policy, the system of equations needs to be solved. Since there are s equations and s unknowns it is possible to solve the system directly. However, if the environment is complex then the curse of dimensionality will make solving the system of equations computationally intense.

Recall the quantity, EVPI, the expected value of perfect information. Now that the Bellman optimality equation has been defined, EVPI can be formally expressed as:

$$EVPI = \sum_i P(H_i)[Q_i^*(s_t, a_t) - \bar{Q}^*(s_t, a_t)] \tag{8}$$

The term, $Q_i^*(s_t, a_t)$ is similarly defined as in equation (7): the value under the optimal action assuming the i^{th} hypothesis about the environemnt. Using equation (8) it is more clear that EVPI is simply the average of optimal expected utilities, assuming that each of the i hypotheses are correct, compared to the optimal expected utility averaged over all hypotheses.: $\bar{Q}^*(s_t, a_t)$. This quantity is also trivial to compute, but requires that $\bar{Q}^*$ be specified.

5 Reinforcement Learning

This discussion so far has decomposed decision analysis into average utility and uncertainity, described how adaptive management incorporates epistemic uncertainty through likelihood updating and aleatoric uncertainty through defining transition probabilities, and introduced the Bellman optimality equation. Since the Bellman optimality equation -a system of equations- can be computationally intensive to solve in complex decision problems, another method needs to be implemented. This motivates the need to understand reinforcement learning (RL). Reinforcement learning is a framework where an agent interacts with its environment in order to obtain the best estimate for the Bellman optimality equation. The following figure displays how an agent interacts with its environment: [9]

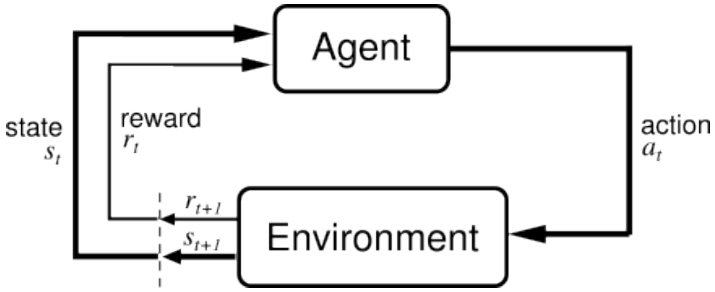


Figure 3: RL Framework

The agent is a construct that updates its knowledge about current rewards and potentially current transition probabilities. The action at time t, a_t, is chosen based on an initialized policy, then the environment is monitored for its subseqent reward and state. Based on the new information, the agent may update its policy, and the Bellman optimality equation, and repeat the cycle until the policy becomes stable. The agent's goal is to maximize the reward it recieves in the long run; that is, the policy is adjusted based on which actions are learned to yield the highest reward. Once this learning becomes stable, then the final estimated policy is achieved.

Two distinguishing factors of RL include: its ability to adjust for actions affecting immediate and subsequent rewards, and its balance of exploration with exploitation of current knowledge in determining the optimal policy. Rewards in RL are re-expressed using: $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$. The parameter γ is referred to as a discount rate and can take values between 0 and 1 inclusively, and k refers to the number of time steps taken into the future. If γ is closer to zero, then the agent prioritizes maximizing immediate rewards as opposed to the rewards within k time steps into the future. However, as γ gets closer to 1, the agent prioritizes maximizing future rewards as opposed to immediate ones. Exploration is incorporated into RL based on how the optimal action is selected. So far, the optimal action has been determined based on the *argmax* function, the action that yields the highest reward. This is known as the greedy algorithm. An ϵ-greedy algorithm chooses an action with the highest reward every time except for ϵ percent of the time. During that ϵ-percent of the time an action is chosen at random. This allows for other actions that may have a delayed impact to be explored. Rather than choosing an action at random during the ϵ-percent of time, a function can be imposed so that actions are chosen based on their

respective action-value. This is referred to as the softmax action, action a_t is chosen with probability:

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}} \tag{9}$$

The term b refers to all of the possible actions in the environment, and τ refers to the temperature. The larger the value of τ, the more likely actions are chosen with almost uniform probability. Through these two characteristics, discount rates and different action-selection strategies, RL provides a flexible framework to make a more informed estimate of the policy- and Bellman optimality equation.

There is an important assumption that needs to be met in order to assure validity of any reinforcement learning solutions. Notice that in Figure 3 the only information that is essential to the agent is the current state and utility. This is known as the Markov Property: the decision must only be dependent on the current environment, and should not be affected by the path the decision process had to take to arrive at the current enviornment. This can be more formally defined as: $P(s_{t+1}, r_{t+1} \mid s_t, a_t, r_t, s_{t-1}, a_{t-1}, ...r_1, s_0, a_0) = P(s_{t+1}, r_{t+1} \mid s_t, a_t)$. If this assumption can be met then several reinforcement learning methods can be implemented to approximate the policy- and Bellman optimality equation.

5.1 A Dynamic Programming Approach

Dynamic programming is a backwards iterative process (Yaesoubi et al. 2011). It is based off the Bellman optimality principle which states that: an optimal policy has the property that whatever the initial state and decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions.[1] In other words, subsections of an optimal policy are still optimal. Dynamic programming incorporates this principle by dividing a policy into subsections. Suppose that a decision-maker knows the expected utility at time *T*. Using the Bellman optimality equation, the expected utility can be estimated for time *T-1*; recall equation (5) if the value of the future time step is known, then the value of the previous time step can be calculated. Based on the Bellman optimality principle, the optimal action determined at time *T-1* constitutes part of the optimal policy. This process can be repeated again for time *T-2*, and so on until a desired time point is reached. The complete collection of actions generates the optimal policy. The construction can be written by the following:

$$t = T : \bar{Q}(a_T, s_T)$$

$$t = T - 1 : \bar{Q}(a_{T-1}, s_{T-1} \mid a_T, s_T))$$
$$\vdots$$
$$t = t : \bar{Q}(a_t, s_t \mid a_{t+1}, s_{t+1})$$

In addition each successive line also assumes that the decision maker knows the transition probabilities and utility of being in a state. Unfortunately, this is a very large assumption since it is possible for decision makers to not know enough information about a disease. This is one drawback of dynamic programming, it can be used only if there is an understanding of the disease dynamics. However, using dynamic programming in combination with adaptive management would allow the decision-maker to update their beliefs of transition probabilites. In the next section, another approach will be introduced where complete information about the environment is unecessary to estimate the Bellman optimality equation. Since the dynamic programming framework starts at the terminal time point, $\bar{Q}(a_T, s_T)$ needs to be defined - where T is the final time point. Fortunately, the terminal action-value quantity can be sometimes be chosen arbitrarily (Q(s,a) = 0 for all s is one example) - most of the time however $\bar{Q}(a_T, s_T)$ is chosen informatively. Policy evaluation is one form of dynamic programming and it yields an approximated action value function for time t. The algorithm is more formally defined below:

1. Initialize Q(s,a) for all s
2. Repeat until $\Delta < \theta$ (some small number)

- Define $\Delta = 0$
- For each state in the enviornment
  - q = Q(s,a)
  - $Q(s,a) = \bar{Q}(s_t, a_t)$
  - $\Delta = max(\Delta, \mid q - Q \mid)$

Once convergence has been reached, that is once $\Delta < \theta$, then $Q(s,a)$ will contain the approximated action-value function. Recall that the deliverable should be an optimal policy, however the algorithm above simply yields an approximate action value function for time *t*. To determine the optimal action, policy iteration needs to be implemented. The algorithm is defined below:

1. Initialize Q(s,a) and $\pi(s)$ - the policy- for all s
2. Perform policy evaluation
3. Perform policy improvement
   - Define *policy stable* = true
   - For each state
     - b = $\pi(s)$
     - $\pi(s) = argmax_{a_t}[\sum_i P(H_i)[R(a_t \mid s_t) + \sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) Q_i(s_{t+1}, a_{t+1})]]$
     - If b $\neq \pi(s)$ then *policy stable* = false
   - If policy is stable then stop, else repeat policy evaluation

Dynamic programming is somewhat limited due to the curse of dimensionality: the number of operations often grows exponentially with the number of state variables. However, in smaller decision problems dynamic programming is preferred to a direct computation.

5.2 Temporal Difference Learning

There are disadvantages with dynamic programming methods: it requires perfect knowledge of the transition probabilities and utilities for a given state, and can be computationally expensive when there are a large number of states and actions. Temporal difference learning is a forward iterative algorithm that also yields an approximated action value function. Using this method, knowledge of the transition probabilties and utilities are not required. Instead, the action value function is learned over time with knowledge of utilities obtained from action-state combinations. SARSA is a temporal difference algorithm that stands for a quintuple of events: $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$. The algorithm is defined below:

- Initialize Q(s,a) arbitrarily
- Repeat for each outbreak simulation
  - Initialize s
  - Choose a from s using policy derived from Q
  - Repeat for each time point in the outbreak
    - Take action a, observe r, s'
    - Choose a' from s' using policy derived from Q
    - $Q(s,a) = Q(s,a) + \alpha[r + \gamma Q(s', a') - Q(s,a)]$
    - s = s', a = a'

– Stop when a terminal state is reached

In the above algorithm, the temporal difference learning occurs through the term: $\gamma Q(s', a') - Q(s, a)$; the new estimate is influenced by the action-value at the next time point. The parameter, α represents a learning rate; it describes the extent to which the error in the old estimate contributes to the new estimate. Just as with dynamic programming the action value function is initialized arbitrarily. The difference in this algorithm is that the action space is searched over according to the policy recommendations ("choose a from s using the policy derived from Q"). Thus in the limit the action space will be entirely searched over. Once the action value function has been approximated, policy iteration can be used to obtain the optimal policy; that is SARSA is performed, policy iteration is performed, and the process is repeated if the policy is not stable. The appeal of temporal difference methods comes from the simplicity in its algorithm, and no requirement of perfect knowledge of the disease dynamics.

6 Future Directions

Adaptive management provides decision makers with optimal strategies based on sequentially collected data while controlling for uncertainty. The reinforcement learning framework provides a means to efficiently estimate the action-value function, and thus the optimal policy. Now that the theory of these two methods have been explained, the original inspiration for understanding optimal decision making in the face of uncertainty, the 2001 foot-and-mouth disease outbreak, can be revisited. The disease dynamics of this outbreak: the rate of transmissibility, the latent period, and the rate of susceptibility have all been previously defined by Dr. Matt Keeling from the University of Warwick. Using that information, he has constructed a means of simulating the foot-and-mouth epidemic. The natural question that arises is: how does adaptive management improve decision-making in the control of a disease outbreak? This question is straightforward to explore; the outbreak can be simulated until a specific time point, the state of the outbreak can be recorded, an optimal decision can be made via a reinforcement learning method, then the outbreak can be monitored to a specific time point so the process can be repeated again. This approach could be compared to a static approach where a specific action is implemented during the entire outbreak period. Learning can also be incorporated in this context by evaluating competing models of risk of infection. Ideally all of the farms in the simulation are recieving an intervention, however it is important that the farms most at risk of infection are prioritized to recieving an intervention. Definition of a risk model will have some degree of uncertainty, thus competing models can be evaluated with accumulation of data. Furthermore, the decision making process could be more spatially explicit. That is, the optimal action could be different across the infected areas. Using the same process of adaptive management intertwined with reinforcement learning, different optimal actions could be chosen for specific clusters of farms. These clusters of farms would again be identified through a risk model, which would be updated using accumulated data. The results of this management strategy could be compared to a global adaptive management-reinforcement learning approach where the same action is implemented everywhere at a given point in time.

In a more theoretical context, an interesting future topic for research would be to compare the efficiency of the different Bellman optimality equation estimation approaches. This discussion has introduced dynamic programming, temporal difference, and Monte Carlo methods (Merl, Johnson, et al, 2009) to approximate the action-value function. A natural question to explore would be: how do these methods compare computationally e.g - rate of convergence) in small decision problems and larger decision problems. In addition, the final policy under each optimization strategy can also be compared among the three methods. Interestingly enough, a thorough comparison of the convergence properties among these methods is yet to be performed.

7 References

1. Bellman, Richard. "On the Theory of Dynamic Programming." Proceedings of the National Academy of Sciences of the United States of America 38.8 (1952): 716-19. RAND. Web. 17 July 2016. <https://www.rand.org/content/dam/rand/pubs/pa

2. Fonnesbeck, Christopher J. "Adaptive Management and the Value of Information: Learning Via Intervention in Epidemiology." PLOS Biology:. N.p., 21 Oct. 2014. Web. 17 July 2016. <http://journals.plos.org/plosbiology/article?id=10.1371%2Fjournal.pbio.1001970>.

3. "Foot-and-Mouth." The Cattle Site. N.p., n.d. Web. 18 May 2016. <http://www.thecattlesite.com/diseaseinfo/243/footandmouth/> .

4. “Foot and Mouth Disease." Animal Health and Welfare: FMD Data Archive. DEFRA, 19 Mar. 2004. Web. 18 May 2016. <http://footandmouth.fera.defra.gov.uk> .

5. Gelman, Andrew. "Bayesian Inference." Bayesian Data Analysis. 3rd ed. Boca Raton: CRC, 2014. 6-7. Print.

6. Krausman, Paul R., and James W. Cain. Wildlife Management and Conservation: Contemporary Principles and Practices. Baltimore: Johns Hopkins UP, 2013. Print.

7. Merl, Daniel, Leah R. Johnson, Robert B. Gramacy, and Marc Mangel. "A Statistical Framework for the Adaptive Management of Epidemiological Interventions." PLoS ONE 4.6 (2009): n. pag. Web.

8. Sit, Vera, and Brenda Taylor. "Decision Analysis: Taking Uncertainties into Account in Forest Resource Management." Statistical Methods for Adaptive Management Studies. Victoria: British Columbia, Ministry of Forests Research Program, 1998. 105-24. Print.

9. Sutton, Richard S., and Andrew G. Barto. Reinforcement Learning: An Introduction. Cambridge, MA: MIT, 1998. Print.

10. “The 2001 Outbreak of Foot and Mouth Disease - National Audit Office (NAO)." The 2001 Outbreak of Foot and Mouth Disease. N.p., n.d. Web. 18 May 2016. <https://www.nao.org.uk/report/the-2001-outbreak-of-foot-and-mouth-disease/> .

11. "UQ - YouQ." UQ - YouQ. Stanford University, n.d. Web. 20 May 2016. <http://web.stanford.edu/group/uq/uq_youq.html>.

12. Williams, Byron K., James D. Nichols, and Michael J. Conroy. Analysis and Management of Animal Populations: Modeling, Estimation, and Decision Making. San Diego: Academic, 2002. Print.

13. Yaesoubi, Reza, and Ted Cohen. "Dynamic Health Policies for Controlling the Spread of Emerging Infections: Influenza as an Example." PLoS ONE 6.9 (2011): n. pag. Web.