# CS 497 SML
# Spring 2003
# Scribe Notes

by Shai Sachs

30 Jan. 2003

## Agenda

(Today's topics in bold)

- Decision Theory
  - Parameter Estimation
  - Utility Theory
  - **Criticisms**
  - **Multi-objective optimality**
- Sequential decision making
  - State transition equation
  - Representations
    - STRIPS
    - decision theoretic
    - graph theoretic

## 1 Criticisms, Continued

### 1.1 Nondeterministic decision making

There are two main criticisms of nondeterministic decision making: first, "worst-case" analysis can yield undesirable results in practice; and second, the same decisions can often be acquired through Bayesian decision making by manipulating the prior distribution.

Consider the rewards in Figure 1. Worst-case analysis causes us to choose $u_1$, although in practice we may want to choose $u_2$ if we know the risk of event

| | $u_1$ | $u_2$ |
|---|---|---|
| $\theta_1$ | 2 | 1000 |
| $\theta_2$ | 2 | 0 |

Figure 1: A scenario in which worst-case decision making might yield undesirable results.

| Bayesian | frequentist |
|---|---|
| A probability is a belief about the outcome of a single trial. | A probability of an event is the limit of the frequency of that event in many repeated trials. |
| subjective | objective, minimalist |
| practical | rigorous, but often useless |

Figure 2: A comparison of the Bayesian and frequentist interpretations of probabilities.

$\theta_2$ is low (equivalently, the probability of $\theta_2$ must be rather high for the expected value of the scenario to favor choosing $u_1$.) Further, we can simulate the result of the nondeterministic decision as a Bayesian decision by correctly assigning prior distributions - for example, by setting $P(\theta_1) \ll P(\theta_2)$.

## 1.2 Bayesian decision making

A common criticism of Bayesian decision making centers around the Bayesian interpretation of probabilities. We compare the Bayesian and frequentist interpretations of probabilities in Figure 2.

While frequentists do not incorporate prior beliefs into decisions, they do incorporate observations. Thus, a frequentist risk function might be:

$$R(\theta, \gamma) = \int_y L(\theta, \gamma(y))p(y|\theta)dy$$

**Prior distributions** One problem with this function is that both $\theta$ and $\gamma$ are unknown. If we are to choose a $\gamma$ which minimizes $R(\theta, \gamma)$, then our choice of $\theta$ might considerably influence the choice of $\gamma$.

A considerable difficulty for Bayesian decision making is determining the prior distributions. One common distribution is the Laplace distribution. Using the principle of insufficient reason, this distribution makes each $\theta$ equally likely.

The Laplace distribution has some justification from information theory. Lacking any information about $\Theta$, we may wish to choose the most "non-informative" prior - that is, the probability distribution which contains the least information.

The entropy contained in a probability distribution over $\Theta$ can be computed using the Shannon Entropy Equations:

$$E = -\sum_{\theta} P(\theta) \log P(\theta) \quad (1)$$

$$E = -\int_{\theta} p(\theta) \log p(\theta) d\theta \quad (2)$$

Equation 1 is for discrete probability mass functions; Equation 2 is for continuous probability density functions.

Using Shannon's Entropy Equations, we can show that the probability distribution which yields the least information (the highest entropy) is that which assigns equal probabilities to all events in $\Theta$ – that is, the Laplace distribution.

**The structure of $\Theta$** We encounter several problems with the Laplace distribution as we consider the structure of $\Theta$.

Suppose $\Theta = \mathbb{R}$. The Laplace distribution assigns 0 probability to any bounded interval of $\mathbb{R}$. This difficulty is mostly mechanical however; the use of generalized probability density functions solves this problem.

Often, we can structure $\Theta$ in arbitrary ways that significantly affect the prior distribution. Suppose we let $\Theta = \{\theta_1, \theta_2\}$, where $\theta_1$ indicates "no precipitation" and $\theta_2$ indicates "some precipitation". The Laplace distribution assigns $P(\theta_1) = P(\theta_2) = \frac{1}{2}$.

Suppose instead we let $\Theta' = \{\theta_1, \theta_2, \theta_3\}$, where $\theta_1$ indicates "no precipitation", $\theta_2$ indicates "rain", and $\theta_3$ indicates "snow". Clearly, $\Theta'$ describes the same set of events as $\Theta$. But in this scenario the Laplace distribution assigns $P(\theta_1) = P(\theta_2) = P(\theta_3) = \frac{1}{3}$. The combined probability of precipitation is $\frac{2}{3}$. Which characterization of nature is correct – $\Theta$ or $\Theta'$?

The following is an interesting practical example of arbitrary choices about the structure of $\Theta$. Suppose we wish to fit a line to a set of points. The equation for the line is $\theta_1 x + \theta_2 y + \theta_3 = 0$. What prior distribution should we choose for $\theta_1$, $\theta_2$ and $\theta_3$? We could choose to "spread the probability" around the unit sphere, by requiring $\theta_1^2 + \theta_2^2 + \theta_3^2 = 1$. However, this choice is entirely arbitrary; we could have also spread the probability around the unit cube, with very different results.

## 2 Multi-Objective Optimality

For now, we concentrate on multiple-objective decisions with no uncertainty. Thus, we have $U$ (the input space), and the loss function $L : U \mapsto \mathbb{R}^d$.

The goal is to find all $u \in U$ such that there is no $u' \in U$ with $L(u') \leq L(u)$. That is, we wish to compute the set of "minimal" inputs $u$, using the partial ordering $\leq$. This set is called the Pareto optimal set of inputs.

Let $L(u) = \langle L_1(u), \ldots, L_d(u) \rangle$. Then we define $L(u') \leq L(u)$. Then we define $L(u') \leq L(u)$ iff $L_i(u') \leq L_i(u)$ for all $i$.

Consider the multi-decision problem depicted in Figure 3. Two robots, indicated by hollow circles, wish to travel along the paths designated. Suppose the

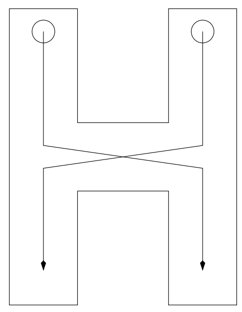

Figure 3: A multi-objective decision problem in which, although there are many conceivable inputs, there are only two Pareto optimal loss values.

path for and speed for each robot is fixed; the only actions possible are starting and stopping at various points in time. Suppose further that the loss function computes the time for each robot to travel along its designated path.

Clearly, many possible inputs are possible. For example, one robot could wait until the other robot has reached its goal before starting to move. Alternately, both robots could move until just before collision, at which point one robot stops and the other continues moving; the stopped robot continues moving once collision is avoided. Nevertheless, there will only be two Pareto optimal loss values for this problem, such as $\langle 4, 6\rangle$ and $\langle 6, 4\rangle$.

One problem with Pareto optimality is that it might yield an "optimal" set which is identical to $U$. For example, consider $U = [0, 1]$ and $L(u) = \langle u, 1 - u\rangle$. It is easy to see that the optimal set for this scenario is just $U$, since whenever one component of $L(u)$ increases, the other decreases by the same amount.

## 2.1 Scalarizing $L$

If we can "scalarize" $L$, then we can find a single optimal value of $L(u)$, rather than many possible optimal values.

We can scalarize $L$ as follows. Choose $\alpha_1, \ldots, \alpha_d \in (0, 1)$. Let $l(u) = \sum_i \alpha_i L_i(u)$. Note that $l(u)$ is just the dot product of $L(u)$ and the $\alpha$ vector.

We can make a multi-objective decision by choosing the $u$ which minimizes $l(u)$. It turns out that this $u$ must also be in the Pareto optimal set. Note that it is possible that more than one $u$ yields the minimum $l(u)$.

We might interpret the $\alpha_i$ as a set of "priorities" over the components of $L$ – higher $\alpha_i$'s are more important, and higher losses in the corresponding components of $L(u)$ should be avoided.