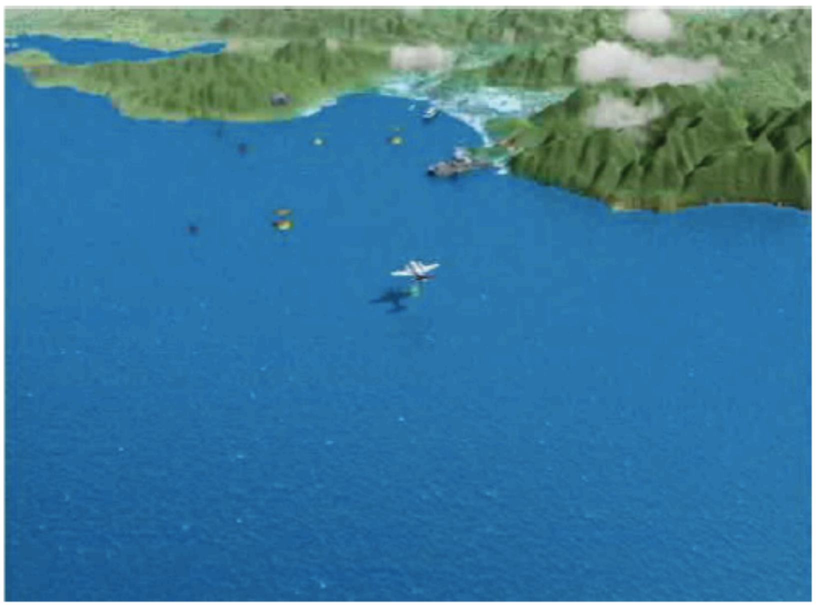
Reinforcement Learning

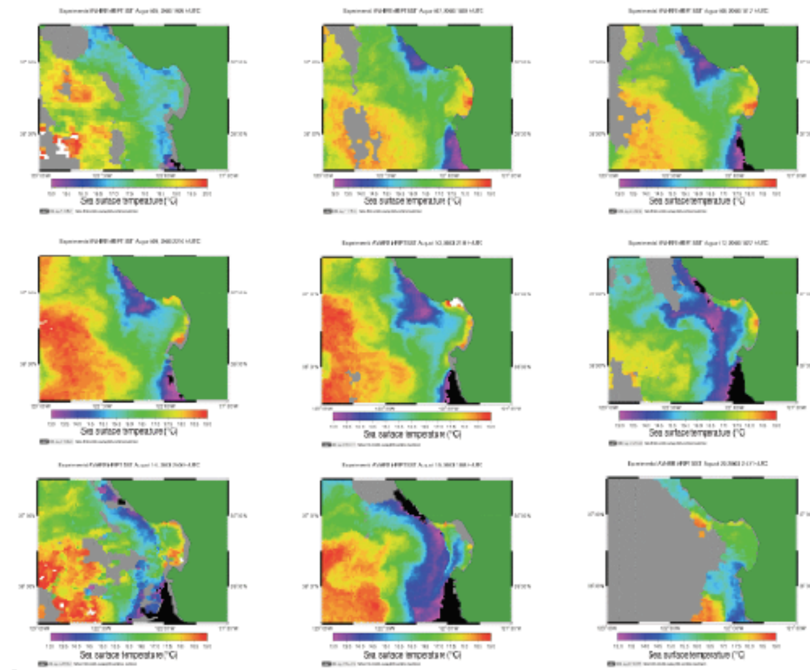
Exploration and Controlled Sensing

Subramanian Ramamoorthy
School of Informatics

21 March, 2017

Example Application: Sampling Spatiotemporal Fields





Satellite Sea Surface Temperature (SST), Monterey Bay, CA, Aug 5-20, 2003

Questions for Ocean Sampling

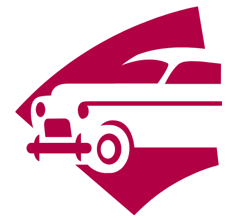
- How to represent the objective that the goal of motion planning is to acquire information which is then used in model learning?

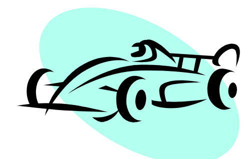
- Concretely, how to decide where and when to sample on the basis of this?

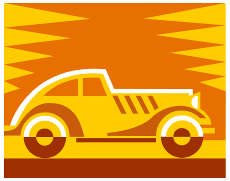
Example Problem: Preference Elicitation

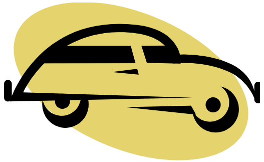
Shopping for a Car:

Luggage Capacity?
Two Door? Cost?
Engine Size?
Color? Options?









Preference Elicitation Problem

... the process of determining a user's preferences/utilities to the extent necessary to make a decision on her behalf

- Issues:
  - preferences vary widely
  - large (multi-attribute) outcome spaces
  - quantitative utilities (the "numbers") difficult to assess
- Preference elicitation can be posed as a POMDP
  - Let us try to formulate the state-action-observation space...

Plan for This Lecture

1. A look (recap) at what Bayesian updating of model parameters achieves

2. Information acquisition problems and the value of information (VoI)

3. Policies based on information gain, e.g., for robots sampling in a navigation setting

Bayesian Updating

Recap of Background

- Learning problem: probabilistic statement of what we believe about parameters that characterise system behaviours
- Focus is on uncertainty about performance:
  - Choice: e.g., of person, technology
  - Design: e.g., policies for running business operations
  - Policy: e.g., when to sell an asset, maintenance decisions
- Beliefs are influenced by observations we make
- Two key ways of thinking about learning problems: frequentist and Bayesian
- Bayesian: start with initial beliefs regarding parameters and combine prior with measurements to compute posteriors

Key Ideas in Bayesian Models

- Begin with a prior distribution over unknown parameter μ
- Any number whose value is unknown is a random variable
- Distribution of the random variable ~ our belief about how likely μ is to take on certain values

$$\mu \sim N(\theta_0, \sigma_0^2)$$

Prior belief

- Bayesian perspective is well suited to information collection
- We always start with some sort of prior knowledge or history
- More important is the conceptual framework that there exists some truth that we are trying to *discover*
- Optimal learning: learn μ as efficiently as possible

Updates for Independent Beliefs

- Consider a random variable, e.g., observation W, normally distributed. We can write its variance and precision as,

$$\sigma_W^2, \beta_W = \frac{1}{\sigma_W^2}$$

- Having seen n observations, we believe mean of μ is θ_n and variance is $1/\beta_n$
- After observing the next measurement we update to,

$$\theta_{n+1} = \frac{\beta_n \theta_n + \beta_W W_{n+1}}{\beta_n + \beta_W}$$

$$\beta_{n+1} = \beta_n + \beta_W$$

Updates for Independent Beliefs

- We could combine these into the more compact form,

$$\theta_{n+1} = (\beta_{n+1})^{-1}(\beta_n \theta_n + \beta_W W_{n+1})$$

- Now, consider the variance of the form,

$$Var_n[\cdot] = Var[\cdot | W_1, W_2, ..., W_n]$$

$$\tilde{\sigma}_n^2 = Var_n[\theta_{n+1} - \theta_n]$$

- This is the variance, given that we have collected n measurements already, so the only random variable at this point is W_{n+1}. Also, think of it as change in variance of θ_n.

Updates for Independent Beliefs

- We could also write θ_{n+1} in a different way by defining the variable,

$$Z = \frac{\theta_{n+1} - \theta_n}{\tilde{\sigma}_n}$$

- This is a random variable only because we have not yet observed W_{n+1}.
- So that we have the update,

$$\theta_{n+1} = \theta_n + \tilde{\sigma}_n Z$$

What Happens to Variance after a Measurement

$$
\begin{aligned}
Var(\mu) &= E[\mu^2] - (E[\mu])^2 \\
&= E(\mu^2) - E[(E[\mu|W])^2] + E[(E[\mu|W])^2] - (E[\mu])^2 \\
&= E[E[\mu^2|W] - (E[\mu|W])^2] + E[(E[\mu|W])^2] - (E[E[\mu|W]])^2 \\
&= E[Var(\mu|W)] + Var[E(\mu|W)]
\end{aligned}
$$

$$E[Var(\mu|W)] = Var(\mu) - Var(E[\mu|W])$$

i.e., variance after measurement will, on average, always be smaller than the original variance. The last term could be zero (if W is irrelevant), but with a sensible signal this is the benefit to measurements.

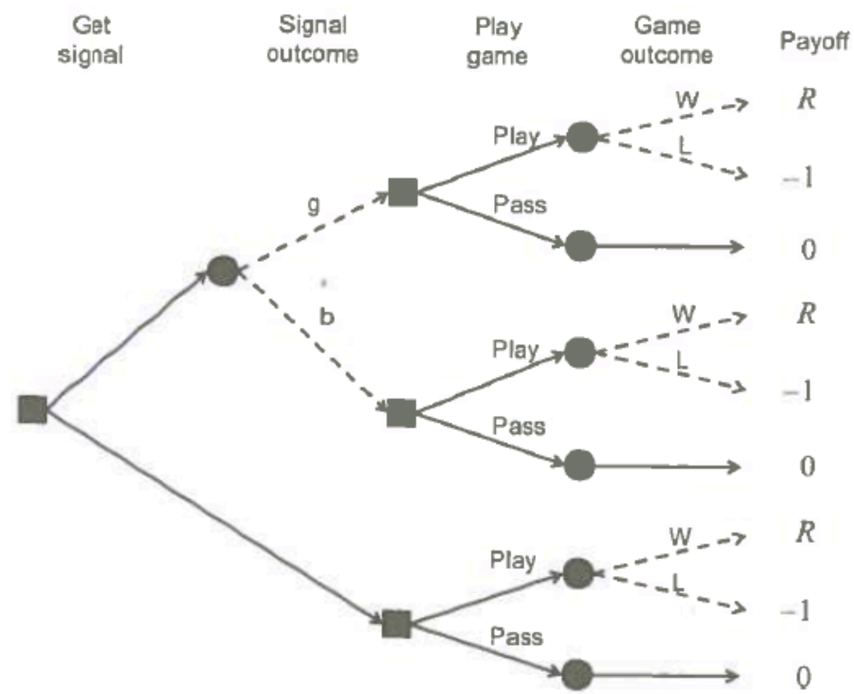
Information Acquisition and VoI

Information Acquisition

- We want to understand the “economics of information”
- Cost of information is highly problem dependent
- Benefits of information can often be captured using models that combine the issues of uncertainty in the context of simple decision problems
- We will look at a simple problem to illustrate key ideas regarding these benefits

Example: Simple Game as a Decision Tree

- We need to decide whether to first acquire a signal that provides information into the probability of winning
- Illustrated in decision tree
- Game has two outcomes:
  - If we win (“W”), we receive reward R
  - If we lose (“L”), we lose -1
  - Lack of information is the information state “N”



Expected Value

- Without any information signal (“N”), probability of winning is known to be p
- Expected value is,

$$E[V|N] = \max\{0, pR - (1-p)\}$$

  – where we assume we will not play if expected value is negative

Remark on notation:
Unlike in our previous discussions where V represented value as in expectation of discounted return, here value will stand for a reward at the end of the game (following convention in litt. on this topic)

Informative Signal

- **Before** we play the game, we have the option of acquiring an information signal S (e.g., purchasing a report or checking information on the internet)
- The signal may be good ("g") or bad ("b")
- We assume that this signal will correctly predict the outcome of this game with probability q, i.e.,

$$P[S = g|W] = P[S = b|L] = q$$

We would like to understand:

- the value of purchasing the signal (elementary information acquisition problem)
- the value of the quality of signal, represented by probability q

Conditional Value

- We first need to understand how the signal changes the expected payoff from the game.
- Conditional value of the game given the signal is,

$$E[V|S=g] = \max\{0, R.P[W|S=g] - P[L|S=g]\}$$

- This equation captures our ability to observe the signal, and then decide whether we want to play the game or not.
- If the signal is bad, expected winnings are,

$$E[V|S=b] = \max\{0, R.P[W|S=b] - P[L|S=b]\}$$

Decision to Acquire

- We next need to find the value of the game given that we have decided to acquire the signal, but before we know its realisation. This is given by,

$$E[V|S] = E[V|S = g]P[S = g] + E[V|S = b]P[S = b]$$

- For this, we need the unconditional probabilities:

$$P[S = g] = P[S = g|W]P[W] + P[S = g|L]P[L]$$
$$P[S = g] = qp + (1 - q)(1 - p)$$
$$P[S = b] = P[S = b|W]P[W] + P[S = b|L]P[L]$$
$$P[S = g] = (1 - q)p + q(1 - p)$$

Conditional Probability of Win/Loss Given the Outcome of Signal

- Use Bayes theorem to write,

$$P[W|S=g] = \frac{P[W]P[S=g|W]}{P[S=g]}$$

$$P[W|S=g] = \frac{pq}{qp+(1-q)(1-p)}$$

- Correspondingly, for the bad signal,

$$P[W|S=b] = \frac{P[W]P[S=b|W]}{P[S=b]}$$

$$P[W|S=g] = \frac{p(1-q)}{(1-q)p+q(1-p)}$$

$$P[L|S=g] = 1 - P[W|S=g], etc.$$

Value of the Signal

- Let S represent the decision to acquire the signal before we know the outcome of the signal.
- Expected value of the game given that we have chosen to acquire the signal is,

$$E[V|S] = E[V|S=g]P[S=g] + E[V|S=b]P[S=b]$$

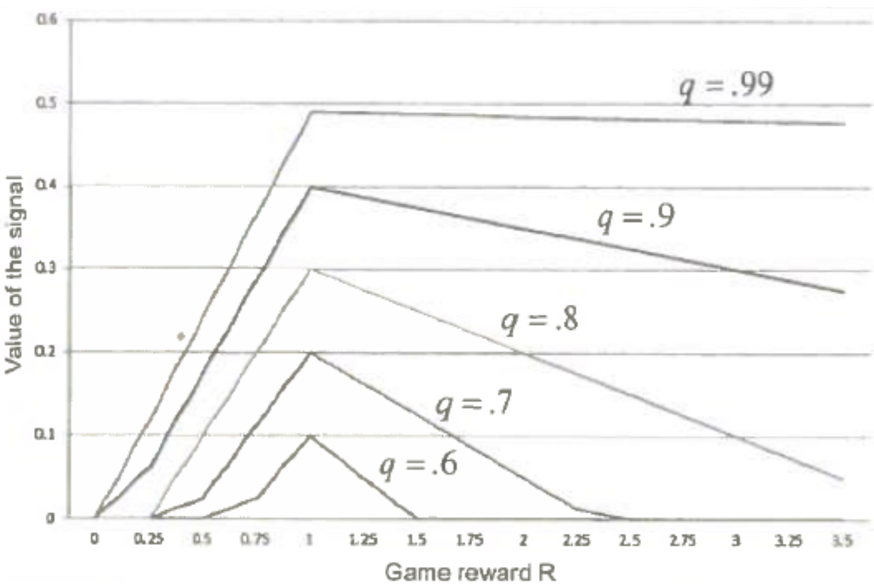
$$E[V|S] = \max\{0, RP[W|S=g] - P[L|S=g]\}(qp + (1-q)(1-p)) +$$
$$\max\{0, RP[W|S=b] - P[L|S=b]\}((1-q)p + q(1-p))$$

$$E[V|S] = \max\{0, R\frac{pq}{qp + (1-q)(1-p)}\}(qp + (1-q)(1-p)) +$$
$$\max\{0, R\frac{p(1-q)}{(1-q)p + q(1-p)} - \frac{q(1-p)}{(1-q)p + q(1-p)}\} +$$
$$((1-q)p + q(1-p))$$

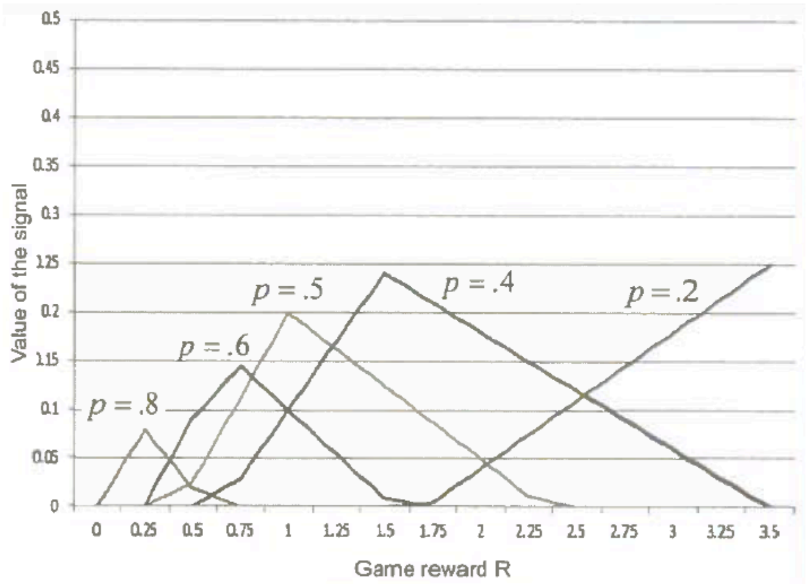
Value of the Signal

- The value of the signal which depends on the game reward R, the probability of winning, p, and the quality of the signal, q,

$$V^s(R,p,q) = E[V|S] - E[V|N]$$



p = 0.5



q = 0.7

Summary of the Simple Example

- We have computed the “value” of a discrete piece of information in a stylized setting.
  - Note that the use of value here, while consistent with our earlier usage, is slightly simpler notationally: the return for a single piece of information does not need a discounted sum

- Next, we turn to a variant where we are allowed to take multiple measurements to increase the precision of the information gained

Towards Marginal Value of Information

- Imagine that we have a choice between doing nothing (with reward 0) and choosing a random reward with mean μ.
- Assume that our prior belief about μ is normally distributed with mean and precision,

$$(\theta_0, \beta_0 = \frac{1}{\sigma_0^2})$$

- Before playing the game, we are allowed to collect a series of measurements, $W_1, W_2, ..., W_n$ (we'll ignore cost for now)
- We assume that W has the unknown mean μ and a known precision β_W

Estimating Reward after n Measurements

- If we choose to make n measurements, the precision of our estimate of the reward would be,

$$\beta_n = \beta_0 + n\beta_W$$

- The updated estimate of our reward (using a Bayesian model) would be,

$$\theta_n = \frac{\beta_0 \theta_0 + n\beta_W \bar{W}_n}{\beta_0 + n\beta_W}$$

$$\bar{W}_n = \frac{1}{n}\sum_{k=1}^{n} W_k$$

- Create a random variable capturing belief about reward
- Use this to make a decision about whether to play the game
- Start with a known identity,

$$Var(\mu) = E[Var(\mu|W_1, ..., W_n)] + Var[E[\mu|W_1, ..., W_n]]$$

where, $$Var(\mu|W_1, ..., W_n) = \frac{1}{\beta^n} = (\beta_0 + n\beta_W)^{-1}$$

$$E[\mu|W_1, ..., W_n] = \theta_n$$

- We can write the change in variance (variance of θ_n given what we knew before we took the n measurements),

$$\tilde{\sigma}^2(n) = Var(\theta_n) = Var(\mu) - E[\frac{1}{\beta_n}]$$

$$\tilde{\sigma}^2(n) = Var(\theta_n) = \frac{1}{\beta_0} - \frac{1}{\beta_n} = \frac{1}{\beta_0} - \frac{1}{\beta_0 + n\beta_W}$$

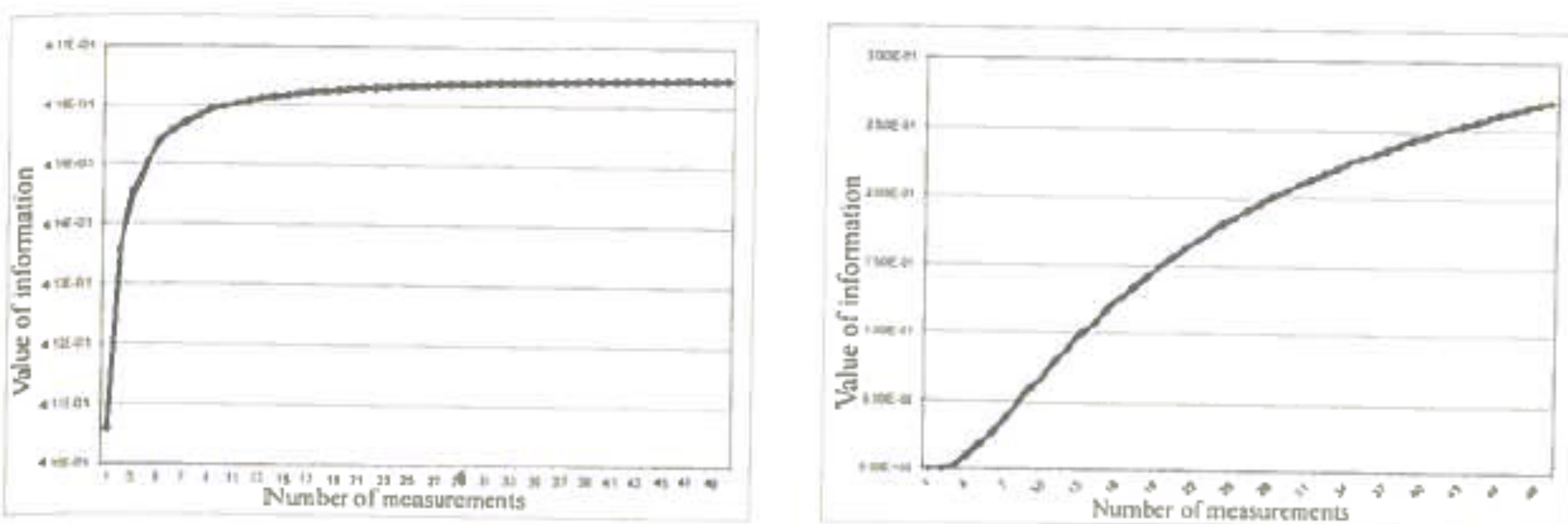
Value of Information

- With Z denoting a standard zero mean –unit variance normal distribution, we can write,

$$\theta_n = \theta_0 + \tilde{\sigma}^2(n)Z$$

- After our n measurements, we are going to choose to play the game if we believe the value of the game is non-zero.
- That value is $V_n = E[\max\{0, \theta_n\}]$
- For each distribution family of interest, one could write down such an expression and expand to get analytical formulation of VoI

Example VoI Curves



(a) (b)

The slope of these curves provide a marginal VoI

Exploration with a Mobile Robot

Exploration Problems

- Exploration: control a mobile robot so as to maximize knowledge about the external world
- Example: robot needs to acquire a map of a static environment. If we represent map as "occupancy grid", exploration is to maximise cumulative information we have about each grid cell
- POMDPs already subsume this function but we need to define an *appropriate payoff function*
- One good choice is information gain:

  Reduction in entropy of a robot's belief as a function of its actions

Exploration Heuristics

- While POMDPs are conceptually useful here, we may not want to use them directly – state/observation space is huge
- We will instead try to derive greedy heuristic based on the notion of *information gain*.
- Limit lookahead to just one exploration action
  - The exploration action could itself involve a sequence of control actions (but logically, it will serve as one exploration action)
  - For instance, select a location to explore anywhere in the map, then go there

Information and Entropy

- The key to exploration is information.
- Entropy of expected information:

$$H_p(x) = -\int p(x) \log p(x) dx \qquad \text{or} \qquad -\sum_x p(x) \log p(x)$$

- Entropy is at its maximum for a uniform distribution, p
- Conditional entropy is the entropy of a conditional distrib.
- In exploration, we seek to minimize the expected entropy of the belief after executing an action
- So, condition on measurement z and control u that define the belief state transition

Conditional Entropy after Action/Observation

- With $B(b,z,u)$ denoting the belief after executing control u and observing z under belief b,
- Conditional entropy of state x' after executing action u and measuring z is given by,

$$H_b(x'|z,u) = -\int B(b,z,u)(x') \log B(b,z,u)(x') dx'$$

- The conditional entropy of the control is,

$$H_b(x'|u) = E_z[H_b(x'|z,u)]$$

$$= \int \int H_b(x'|z,u) p(z|x') p(x'|u,x) b(x) dz dx' dx$$

Greedy Techniques

- Expected information gain lets us phrase exploration as a decision theoretic problem.

- Information Gain is

$$
\begin{aligned}
I_b(u) &= H_p(x) - H_b(x'|u) \\
&= H_p(x) - E_z[H_b(x'|z,u)]
\end{aligned}
$$

Greedy Techniques

- If $r(x,u)$ is the cost of applying control u in state x (treating cost as negative numbers), then optimal greedy exploration for the belief b maximizes difference between information gain and cost,

$$\pi(b) = \arg\max_u \alpha(H_p(x) - E_z[H_b(x'|z,u)]) + \int r(x,u)b(x)dx$$

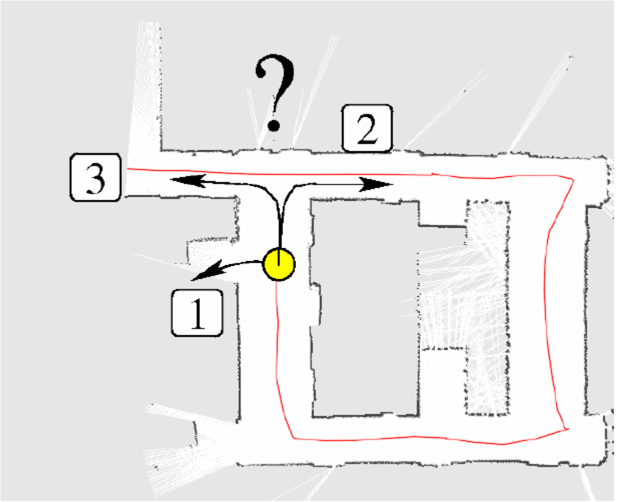
Expected information gain
(Original entropy – Cond. Entropy)

Expected cost

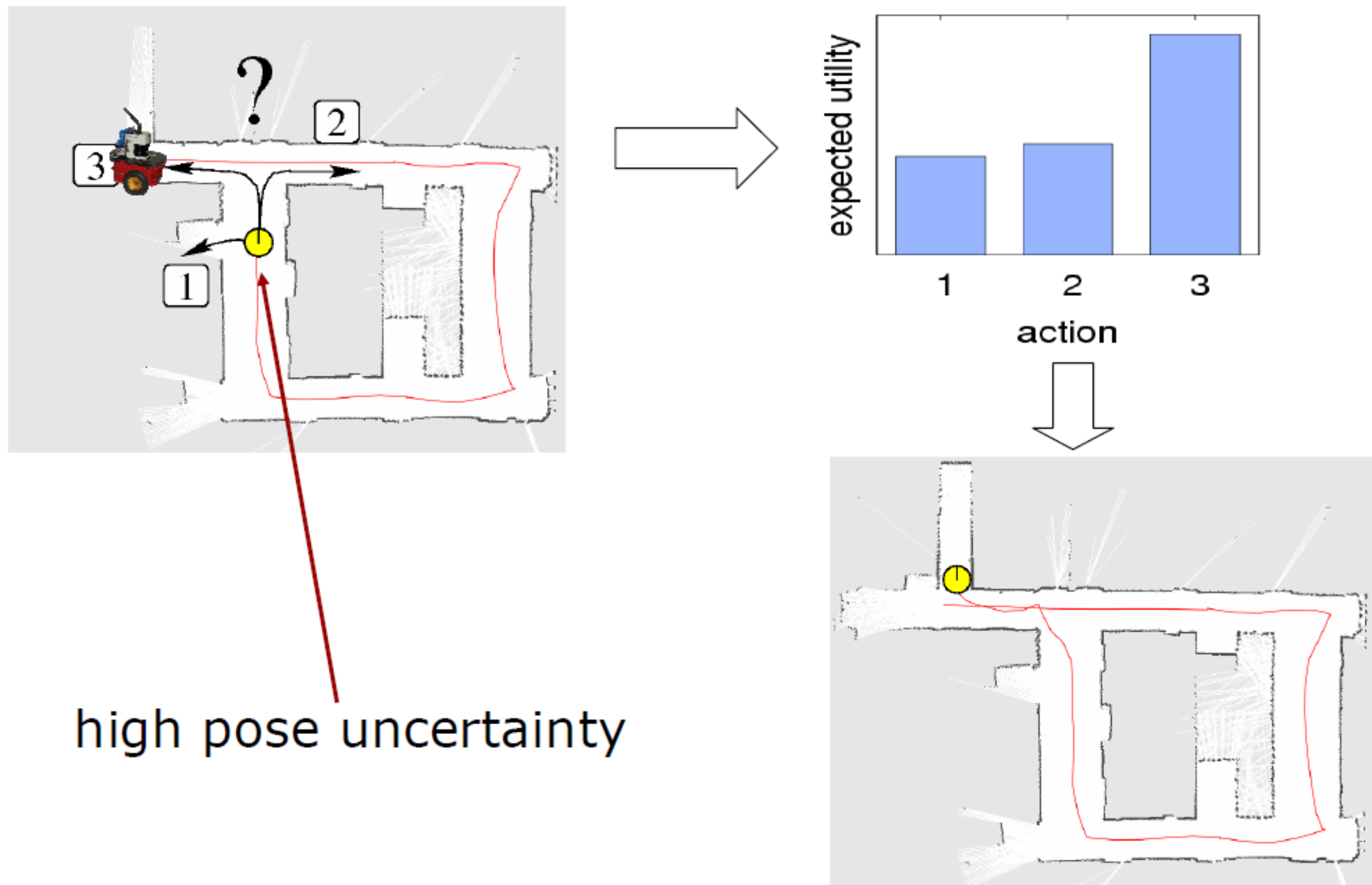
Example: Combining Exploration and Mapping

- By reasoning about control, the mapping process can be made much more effective

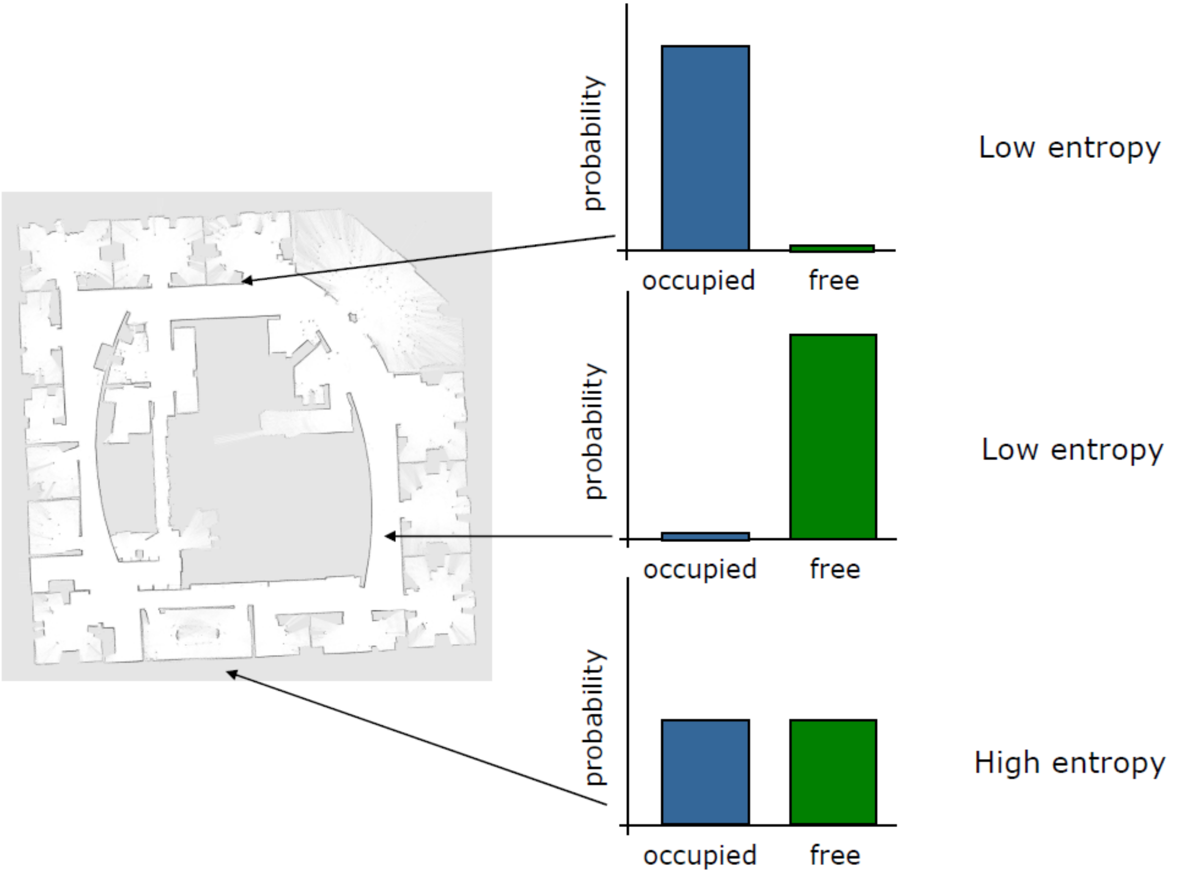
- Question: Where to move next in a map?



Exploration Problem: Visually



Map Entropy



The overall entropy is the sum of the individual entropy values