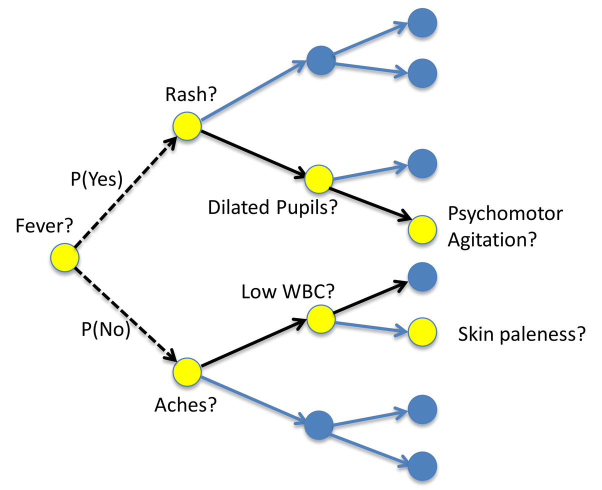
Near-Optimal Bayesian Active Learning with Noisy Observations

Caltech

Daniel Golovin, Andreas Krause, Debajyoti Ray

Bayesian Active Learning

- How should we perform experiments to determine the most accurate scientific theory among competing candidates?
- How should we decide among expensive medical procedures to accurately determine a patient's condition?
- How should we select which labels to obtain in order to determine the hypothesis that minimizes generalization error?



We have to sequentially select among a set of noisy, expensive observations in order to determine which hypothesis is most accurate.

Bayesian Active Learning problem:

Distinguish among a set of hypotheses $\{h_1, \cdots, h_n\}$ by performing tests from a set $T = \{1, \cdots, N\}$ of possible tests. Running test t incurs cost $c(t)$ and produces an outcome in $\{1, \ldots, \ell\}$

We have a prior distribution *P* modeling assumptions on the joint probability $P(H, X_1, \cdots, X_N)$ over the hypotheses and test outcomes.

Noise-free Bayesian Active Learning:

Suppose that $P(X_t \mid H)$ is deterministic (noise-free)

Each test rules out a set of hypotheses, based on its outcome.

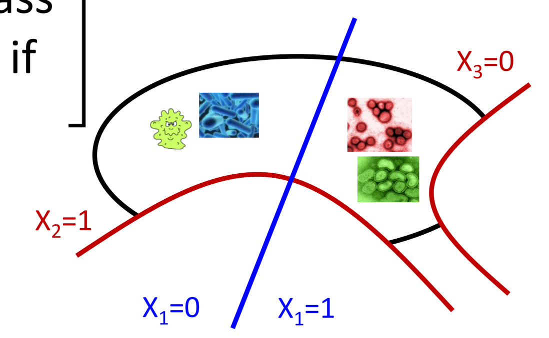
How should we test to rule out all incorrect hypotheses?

$$\Delta_{GBS}(s \mid x_A) = \mathbb{E}\left[\begin{array}{l}\text{probability mass}\\ \text{ruled out by } s \text{ if}\\ \text{we know } x_A\end{array}\right]$$

Generalized Binary Search (GBS):

Greedily maximize Δ_{GBS}

(equivalent to maximizing info-gain.)



Adaptive Submodularity

Key insight: GBS is adaptive submodular

$$\Delta_{GBS}(s \mid x_A) \geq \Delta_{GBS}(s \mid x_B) \text{ whenever } x_A \preceq x_B$$

x_B includes all observations in x_A and possibly more.

Adaptive submodularity [Golovin & Krause, COLT 2010] generalizes submodularity to the adaptive setting.

Adaptive-Greedy is a $\boxed{(\ln(1/p_{\min}) + 1)}$ *approximation*

Results require that tests are exact (no noise)!

Bayesian Active Learning with Noisy Observations:

In practice, observations are noisy. Results for noise free case do not generalize.

Key Problem: Tests no longer eliminate hypotheses (only make them less likely)

Suppose all tests are run, see x_V, best we can do is maximize expected utility:

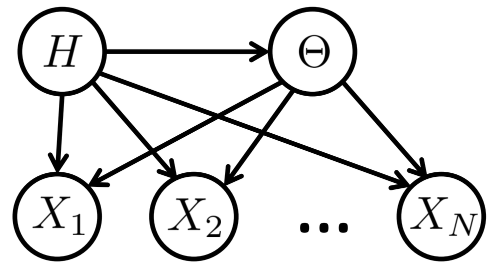
$$a^* = \arg\max_a \sum_y P(y \mid \mathbf{x}_V) U(a, y)$$

How should we cheaply test to guarantee that we choose a^*?

Existing approaches:
- Generalized binary search?
- Maximize information gain?
- Maximize value of information?

Not adaptive submodular in the noisy setting!

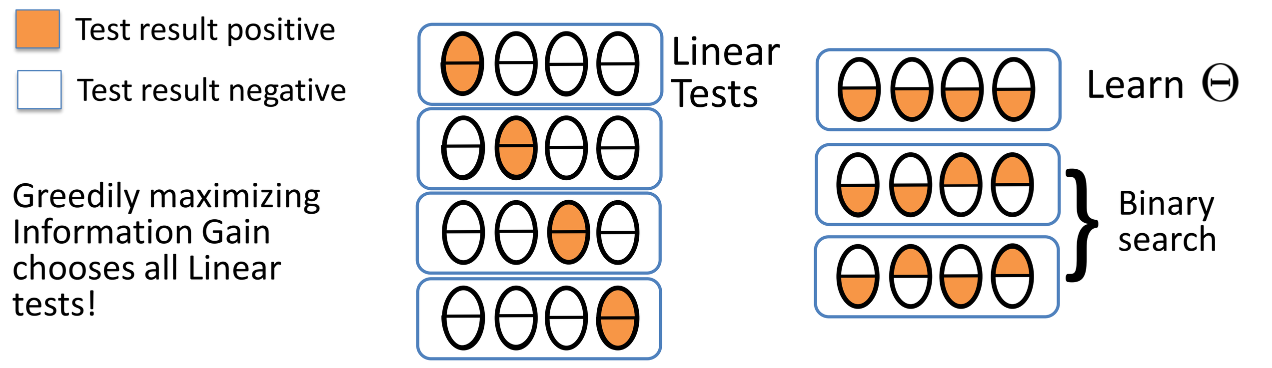
Our noise model:



With noisy observations, $P(X_1, \ldots, X_n \mid H)$ is not deterministic. The noise is modeled with a random variable Θ, so that $P(X_1, \ldots, X_n \mid H, \Theta)$ is deterministic.

Greedily maximizing Information Gain is not adaptive submodular in the noisy case:

Example: H is a random oval, and $\Theta \in \{\text{up, down}\}$



Greedily maximizing Information Gain chooses all Linear tests!

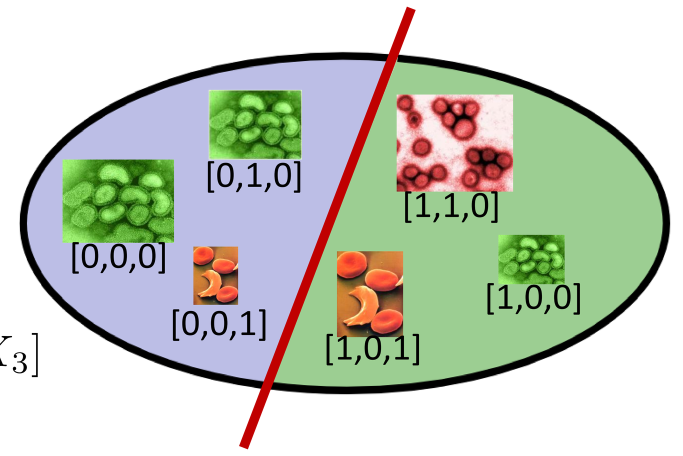
Theorem: All previous approaches pay $\Omega\left(\frac{n}{\log n}\right)$ times the optimum with noisy observations, in the worst case, even when Θ has constant–sized support, and $p_{\min} = \Theta(1/n)$

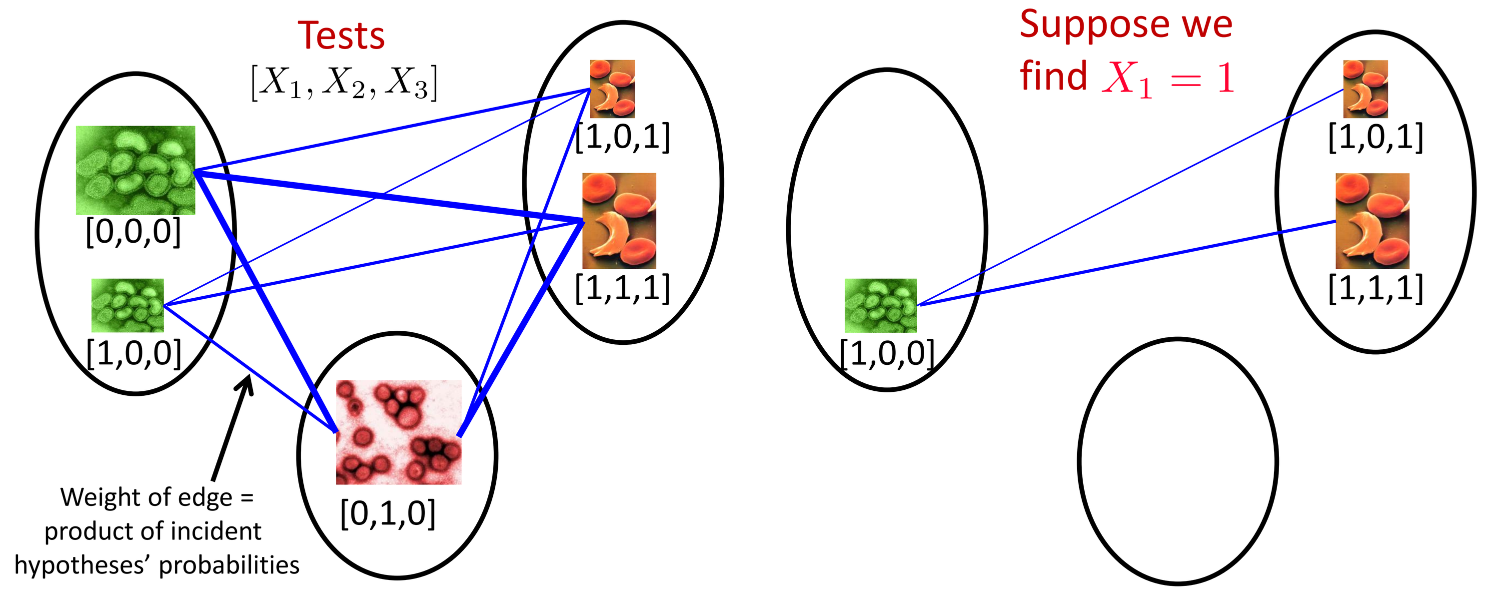
Equivalence-Class Edge-Cutting (EC²)

Strategy: Reduce noisy problem to noiseless problem

Key Idea: Make test outcomes part of the hypothesis

$$\mathbf{X}_V = [X_1, X_2, X_3]$$





*The expected conditional marginal benefit for test **t** upon observations* x_A:

$$\Delta_{EC}(t \mid x_A) = \mathbb{E}[\text{weight of edges eliminated}]$$

Theorem: For the adaptive greedy policy π_{EC} implemented by EC², it holds that

$$c(\pi_{EC}) \leq (2\ln(1/P_{\min}) + 1)\, c(\pi^*)$$

where $P_{\min} := \min\{P(\mathbf{x}_V) : P(\mathbf{x}_V) > 0\}$ is the minimum prior probability of any outcome vector, and π^* is the optimal policy.

Reducing Bayesian active learning to Equivalence Class Determination may in some instances result in exponentially-large equivalence classes, which makes running EC² challenging. We can use rejection sampling.

Alternatively we develop the **Eff**icient **E**dge **C**utting appro**X**imate objec**tive** algorithm that approximates the EC² objective function:

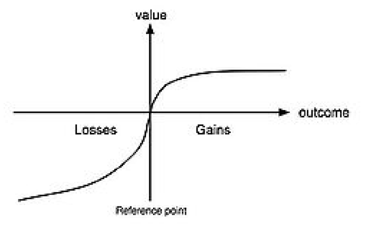
The EffECXtive objective function

$$\Delta_{Eff}(t \mid x_A) := \sum_x P(X_t = x \mid x_A)\Big(\sum_i P(h_i \mid x_A, X_t = x)^2\Big) - \sum_i P(h_i \mid x_A)^2$$

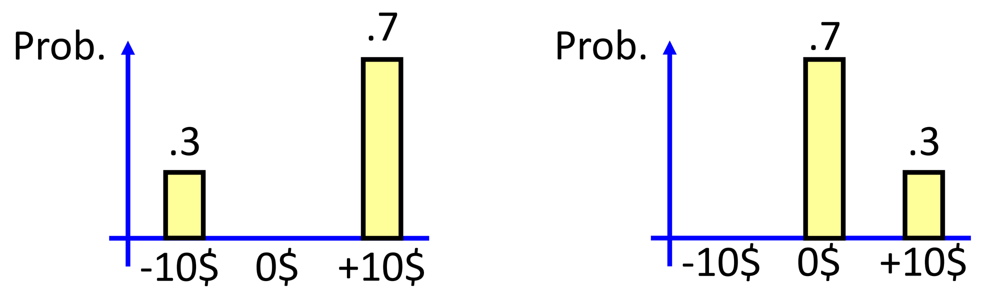
Adaptive experimental design in behavioural economics

We behaviorally test theories of decision-making under uncertainty where the tests are generated dynamically using the EffECXtive algorithm.

- Prospect Theory $U(x,p) = \sum_{x<0} (-\lambda) p_i |x_i|^\rho + \sum_{x>0} p_i |x_i|^\rho$
- Mean-Variance-Skewness $U(x,p) = \mu - \alpha_\sigma \sigma^2 + \alpha_\nu \nu$
- Expected Value $U(x,p) = \sum_x p_i x_i$
- Constant Relative Risk Aversion $U(x,p) = \sum_x p_i (x_i^{1-\rho})/(1-\rho)$



What would you prefer?

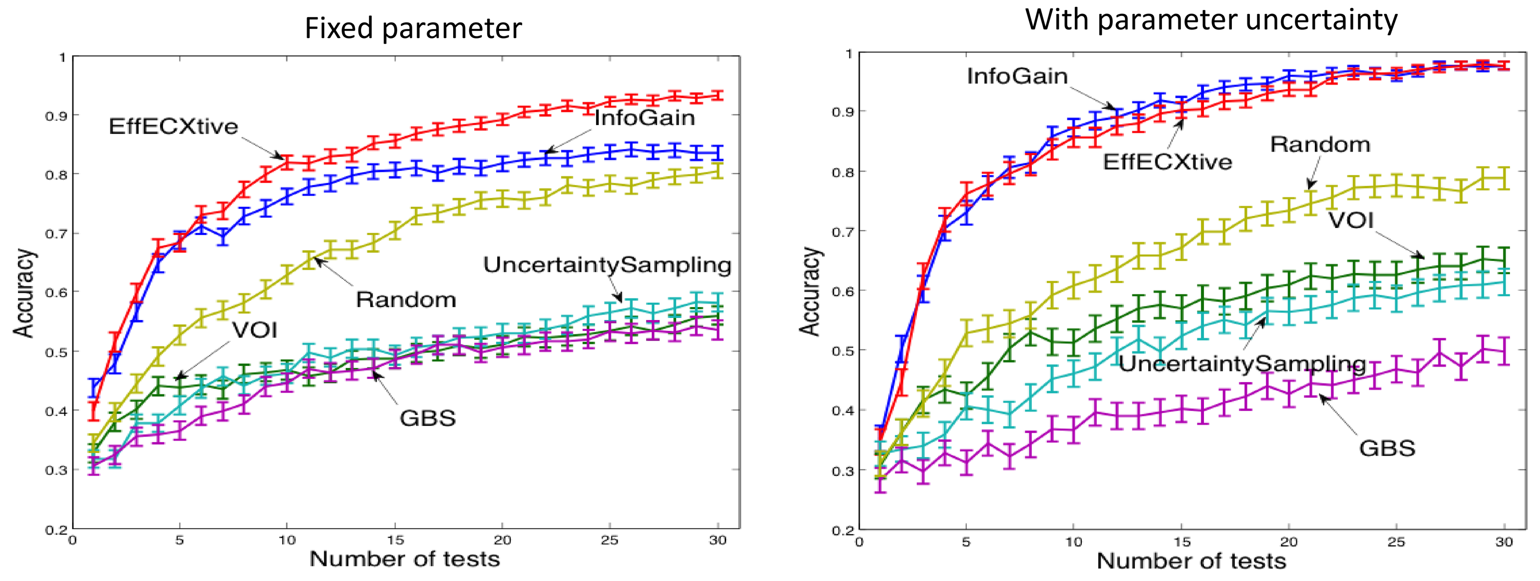


Experimental design:

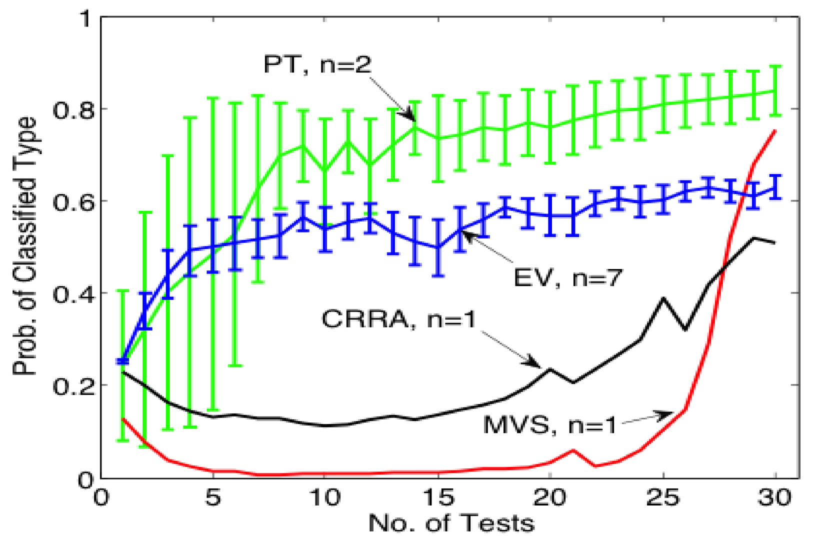
Choice between 2 lotteries. Each lottery has a loss, neutral and gain outcome with varying probabilities.

We varied the probabilities for the outcomes.

Ground truth analysis: We randomly generate a *true* hypothesis and parameter. We pick the MAP hypothesis after 30 tests and check if it corresponds to truth. We repeat this for 1000 trials.



EffECXtive outperforms InfoGain when the hypotheses are identifiable, and performs as well as InfoGain when there is parameter uncertainty, which violates the identifiability assumption of EC².



We tested 11 human subjects using Caltech IRB protocols.

Most subjects (n=7) were classified as Expected Value types.

Some subjects (n=2) exhibited risk aversion and loss aversion and were classified as Prospect Theory types.

One subject violated stochastic dominance and behaviour was best classified using Mean-Variance-Skewness theory.