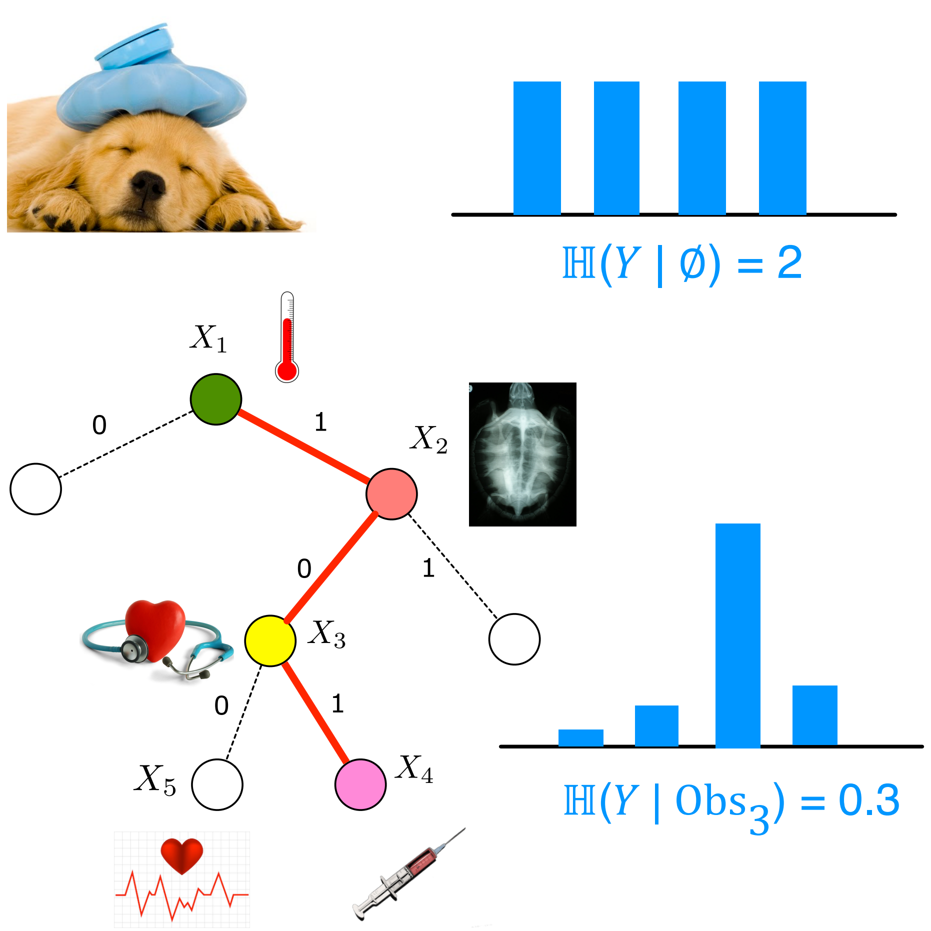
Sequential Information Maximization: *When is Greedy Near-optimal?*

Yuxin Chen†, S. Hamed Hassani†, Amin Karbasi‡, Andreas Krause†

ETH zürich

Yale University

Motivating Application



Applications

- Active learning [MacKay, 1992; Settles, 2012]
- Experimental design [Lindley, 1956; Fedorov, 1972]
- Evaluation of (stochastic) Boolean functions [Kaplan et al., 2005]
- Channel coding with feedback [Horstein, 1963]
- Active hypothesis testing [Chernoff, 1959]

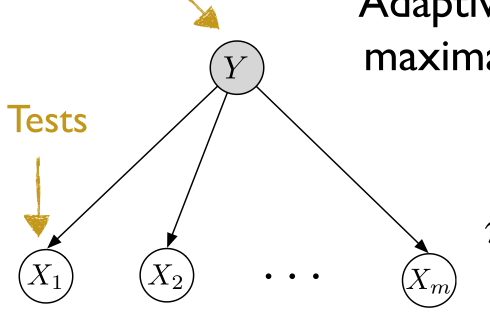
…

The Sequential Information Maximization Problem

Hypotheses: $Y \in \{y_1, \ldots, y_n\}$

Problem

Adaptively find k tests that are maximally informative about Y



$$\pi^* = \operatorname*{argmax}_{\pi \in \Pi_{[k]}} \mathbb{I}(\pi; Y)$$

set of policies of length k

Mutual information between π and Y, i.e., the expected reduction in entropy of Y, given that we have run π.

The Greedy Algorithm

At step ℓ, pick

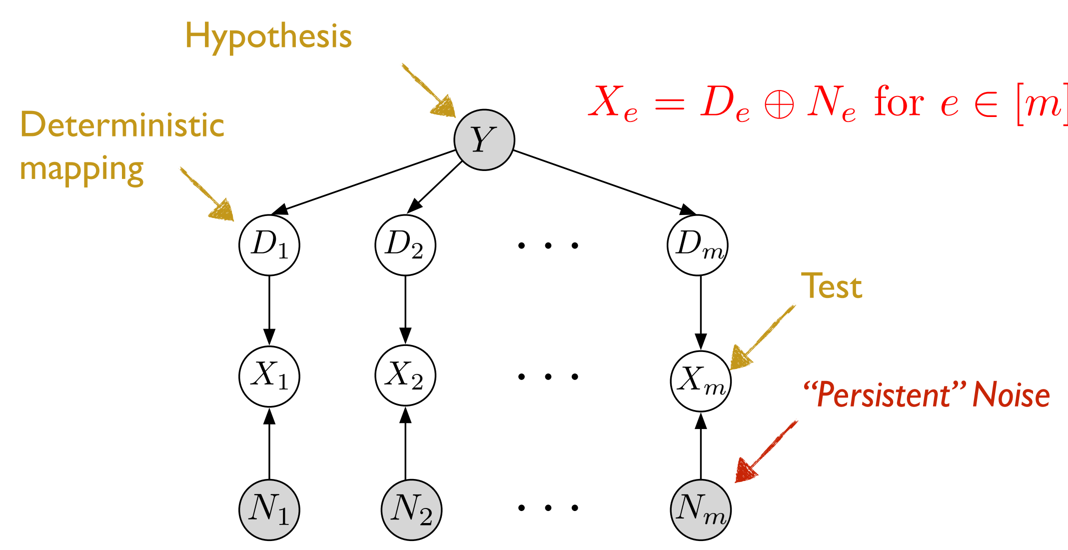
$$e_\ell = \operatorname*{argmax}_{e \in [m]} \mathbb{I}(X_e; Y \mid \text{outcomes of } e_1, \ldots, e_\ell).$$

The most informative selection policy has been used since 1950's [Lindley, 1956]

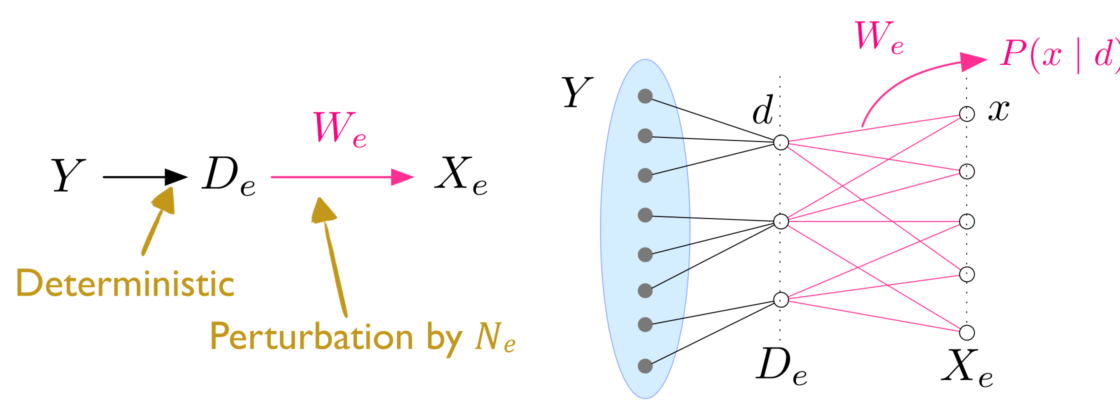
In the non-adaptive setting, Greedy is near-optimal [Krause and Guestrin, 2005]	In the noiseless setting, Greedy is near-optimal [Dasgupta, 2005; Golovin and Krause 2011]	Tests are noisy, but can be repeated with i.i.d. outcome — Reduction to Noiseless Case [Nowak 2009]

In this paper, we present the first rigorous analysis of the most informative selection policy in the persistent noise setting.

An Equivalent Representation



Channel Induced by Noise

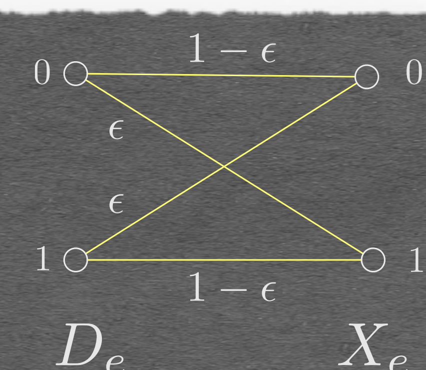


Example

$D_e \in \{0,1\}$ deterministic given Y

$N_e \in \{0,1\}$ with $\Pr(N_e = 1) = \epsilon$

$X_e \in \{0,1\}$ $X_e = D_e \oplus N_e$



Separability of a Channel

$$S(W) = \left(\min_{d, d' \in D : d \neq d'} \left| p(\cdot \mid d) - p(\cdot \mid d') \right|_{\text{TV}} \right)^2.$$

Indicator of how much the channel can differentiate between d and d′.

Example (GBS with persistent noise)

$$S(\text{BSC}) = \left(\frac{1}{2} \sum_{x \in \{0,1\}} \left| p(X = x \mid D = 0) - p(X = x \mid D = 1) \right| \right)^2 = (1 - 2\epsilon)^2$$

$$S_{\min} = \min_{e \in [m]} S(W_e)$$

Minimum value of separability over all the channels W_e

Main Results

For any $\delta > 0$, it holds that

$$\mathbb{I}\left(\pi_{\text{Greedy}[k']}; Y\right) \geq \left(\mathbb{I}\left(\pi_{\text{OPT}[k]}; Y\right) - \delta\right)\left(1 - \exp\left(-\frac{k' S_{\min}}{7k \max\{\log n, \log \frac{1}{\delta}\}}\right)\right)$$

Example (GBS with persistent noise)

$$\mathbb{I}\left(\pi_{\text{Greedy}[k']}; Y\right) \geq \left(\mathbb{I}\left(\pi_{\text{OPT}[k]}; Y\right) - \delta\right)\left(1 - \exp\left(-\frac{k'(1-2\epsilon)^2}{7k \max\{\log n, \log \frac{1}{\delta}\}}\right)\right)$$

Key Lemma

Consider any adaptive policy π which chooses k tests. We must have:

$$\max_{e \in [m]} \mathbb{I}(X_e; Y) \geq \frac{\mathbb{I}(\pi; Y)}{k \cdot \frac{7}{S_{\min}} \cdot \max\left\{\log n, \log \frac{1}{\mathbb{I}(\pi; Y)}\right\}}.$$

Gain of π in k steps

Gain of greedy at a single step

constant

# of hypotheses

For *adaptive monotone* and *adaptive submodular* objective function f:

$$\max_{e \in [m]} \text{Gain}_f(X_e \mid \text{past observation}) \geq \frac{\text{Gain}_f(\pi \mid \text{past observation})}{k}.$$

From Key Lemma to Main Results

Gain of π in k steps

$$\mathbb{I}\left(\pi_{\text{Greedy}[\ell+1]}; Y\right) - \mathbb{I}\left(\pi_{\text{Greedy}[\ell]}; Y\right) \geq \text{const} \cdot \left(\mathbb{I}(\pi; Y) - \delta - \mathbb{I}\left(\pi_{\text{Greedy}[\ell]}; Y\right)\right)$$

Gain of greedy at a single step

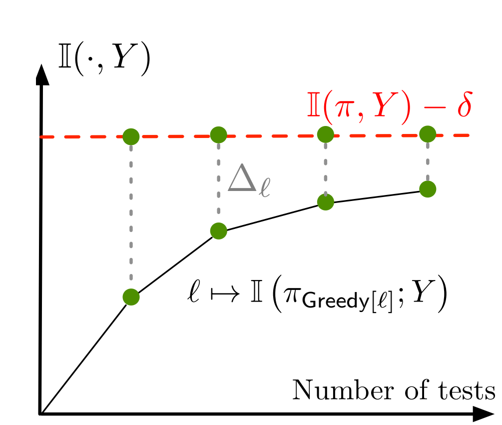
$$\frac{S_{\min}}{7k \max\{\log n, \log \frac{1}{\delta}\}}$$

$$\Delta_\ell := \mathbb{I}(\pi; Y) - \delta - \mathbb{I}\left(\pi_{\text{Greedy}[\ell]}; Y\right)$$

$$\Delta_\ell - \Delta_{\ell+1} \geq \text{const} \cdot \Delta_\ell$$

$$\Delta_{\ell+1} \leq (1 - \text{const})\, \Delta_\ell$$

$$\Delta_{[k]} \leq (1 - \text{const})^k\, \Delta_0 \leq e^{-k \cdot \text{const}} \Delta_0$$



$$\mathbb{I}\left(\pi_{\text{Greedy}[k]}; Y\right) \geq \left(\mathbb{I}(\pi; Y) - \delta\right)\left(1 - e^{-k \cdot \text{const}}\right)$$

Key Lemma — Proof Outlines

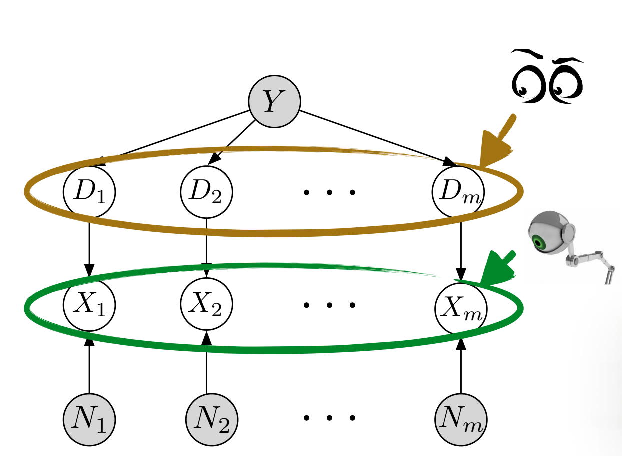
Step 1. If $\mathbb{I}(X_e; Y)$ is small, then the probability of the most likely outcome of test e (i.e., $p_{e,\max} = \max_d \Pr(D_e = d_e)$) is large.

$Y \rightarrow D_e \xrightarrow{W_e} X_e$ $\quad \mathbb{I}(X_e; Y) = \mathbb{I}(X_e; D_e)$

deterministic mapping — noise channel (perturbation by N_e)

Intuitively, smaller $\mathbb{I}(X_e; D_e) \Rightarrow$ more skewed distribution over D_e.

Step 2. Consider an oracle G sitting beside the system π, and observing its actions. For each test e that π picks, G also observes the deterministic (i.e., the *"noise-free"* version) outcome d_e. One can show that: $\mathbb{I}(\pi; Y) \leq \mathbb{I}(G; Y)$.

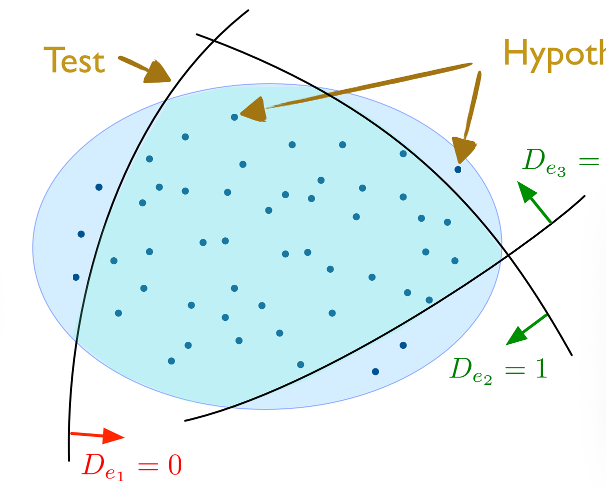


The oracle G observes what π observes (i.e., X_e's), as well as the true outcome (i.e., D_e's)

π observes noise-corrupted test outcomes (i.e., X_e's)

We expect that G has a better idea about Y than π has.

Step 3. If for all test $e \in [m]$, $p_{e,\max}$ is large, then $\mathbb{I}(G; Y)$ is small.



"Good event": event with most likely outcome.

The conditioning entropy of Y given that *"good event" is realized* is high

$\Rightarrow \mathbb{I}(G; Y)$ is low.

$\Pr(D_{e_1} = 0), \Pr(D_{e_2} = 1), \Pr(D_{e_3} = 1)$ large $\Rightarrow \Pr(D_{e_1} = 0, D_{e_2} = 1, D_{e_3} = 1)$ large. (Union B.)

Combining Step 1–3, $\mathbb{I}(X_e; Y)$ is small $\Rightarrow \mathbb{I}(\pi; Y)$ is small.

The Necessity of $S_{\min}$

If $S_{\min}$ is sufficiently large, one can find a "*Treasure Hunt*" example, where the ratio between the gain of greedy and the smarter policy is at most $c\, S_{\min}$, where c is some constant.