Near-optimal Bayesian Active Learning with Correlated and Noisy Tests

Yuxin Chen, S. Hamed Hassani, Andreas Krause

ETH zürich | Learning & Adaptive Systems

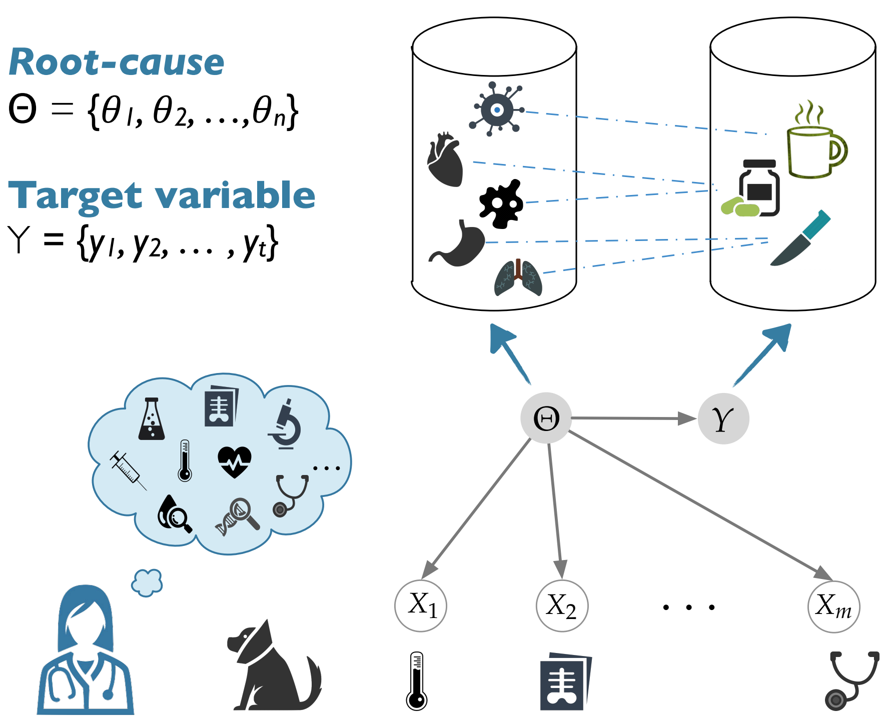
1 Problem Setup

Applications Sequential experimental design; Interactive troubleshooting; Active testing; Active learning …

Example Medical diagnosis

Root-cause $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$

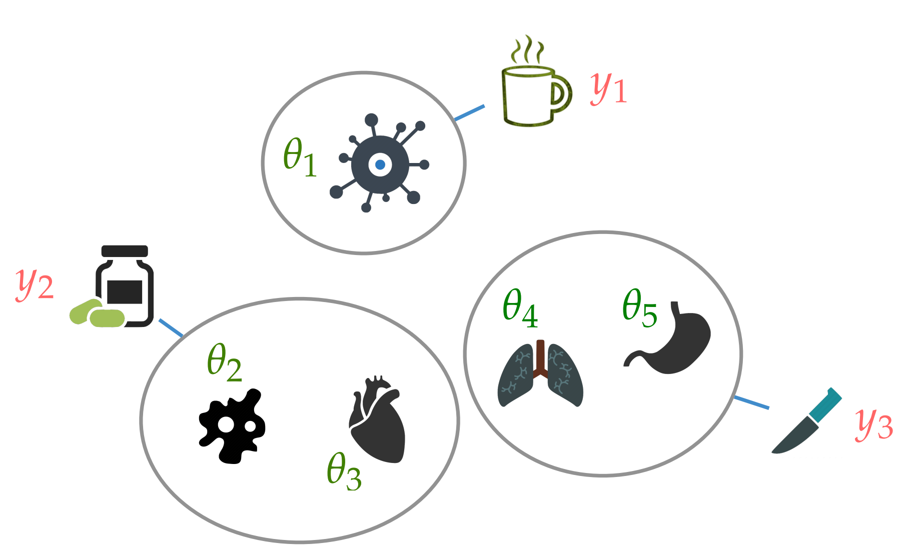
Target variable $Y = \{y_1, y_2, \dots, y_t\}$



Tests $T = \{1, \dots m\}$, each with unit cost

Test outcomes $X_{1:m}$ are correlated with Θ

(Noisy) Bayesian Active Learning



Def. After performing a set of tests A, and observing outcomes $\mathbf{x}_A$, the error probability of the MAP estimator of Y is

$$p_{\text{ERR}}(\mathbf{x}_\mathcal{A}) = 1 - \max_{y \in \mathcal{Y}} \mathbb{P}[y \mid \mathbf{x}_\mathcal{A}]$$

Given Prior $P(\Theta)$

Conditional probabilities $P(X \mid \Theta) \in [0,1]$

Mapping r from Θ to Y: $r(\theta) = y$

Want Min-cost policy π^*, with expected prediction error at most δ:

$$\pi^* \in \arg\min_\pi \operatorname{cost}(\pi),$$
$$\text{s.t. } \mathbb{E}_{x_\mathcal{T}}[p_{\text{ERR}}(\mathcal{S}(\pi, x_\mathcal{T}))] \le \delta.$$

Observations seen by policy π, if the tests outcomes are realized as $\mathbf{x}_T$

Existing Approaches

Greedy Heuristics

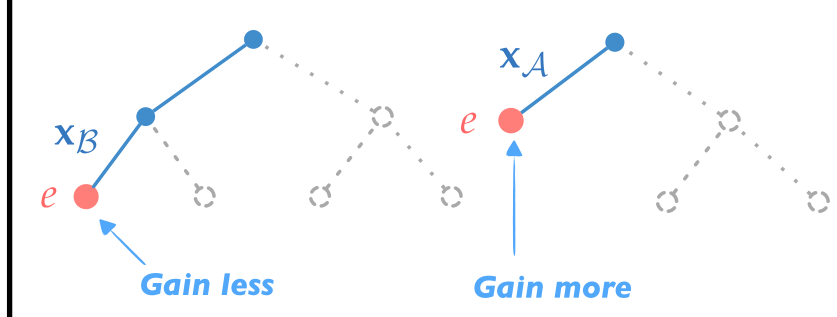
- Uncertainty sampling
- Maximize information gain [e.g., Lindley'56]
- Maximize value of information [e.g., Howard'60]

Thm. All these approaches may require exponentially more tests than the optimum

The Equivalence Class Determination Problem

- Assuming tests' outcomes are noise-free: $P(x \mid \theta) \in \{0,1\}$
- ***Equivalence class edge cutting*** [Golovin & Krause, 2010; …]
- Greedy is near-optimal: $\operatorname{cost}(\pi) = (2 \ln(1/p_{\min}) + 1) \cdot \operatorname{cost}(\pi^*)$

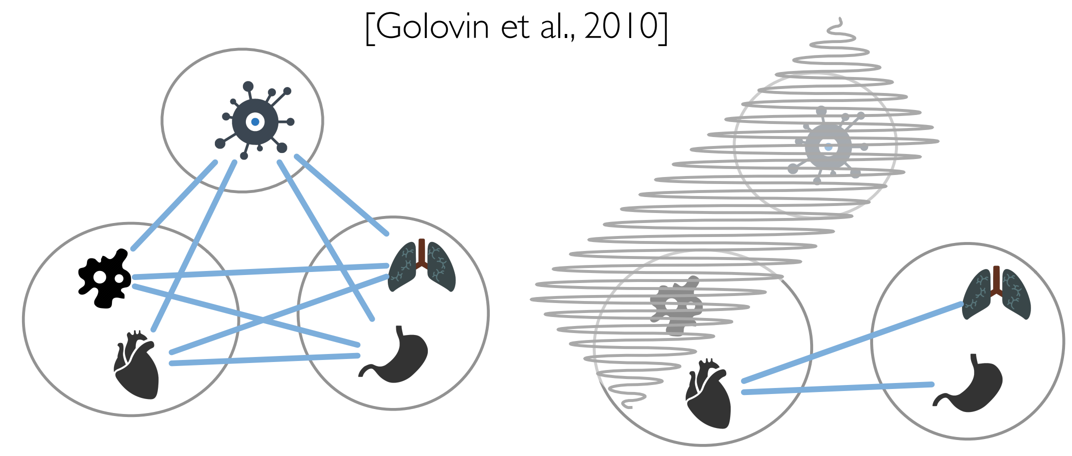
2 Adaptive Submodularity [Golovin & Krause, 2010]



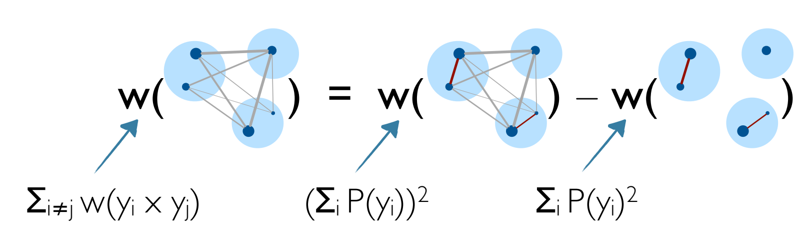
Playing an action earlier only increases its expected marginal benefit.

EC²: Equivalence Class Edge Cutting

[Golovin et al., 2010]

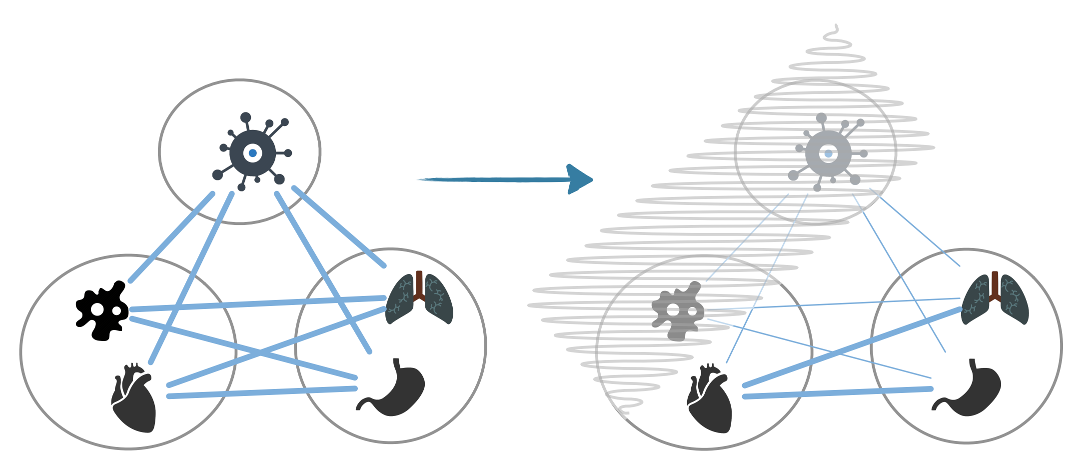


Equivalence Class Edge Cutting (EC²) [Golovin et al., 2010]



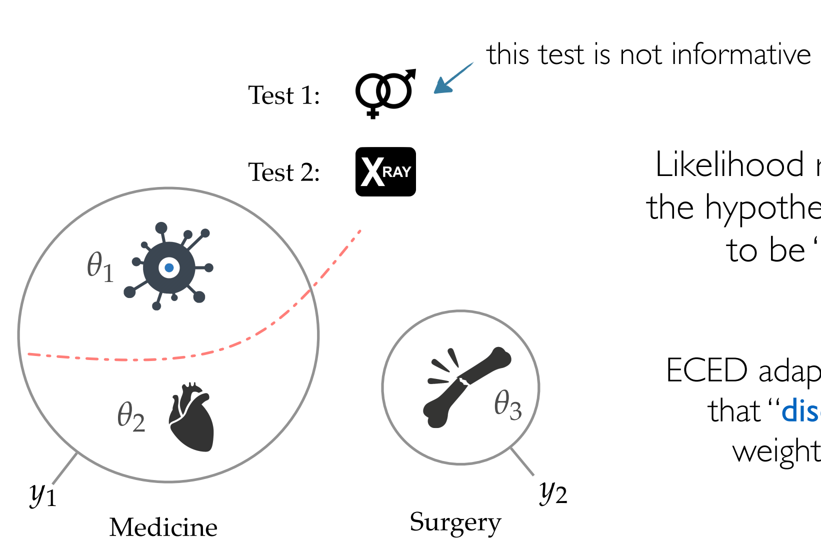
Extension to the Noisy Setting

[This paper — ECED: Equivalence Class Edge Discounting]



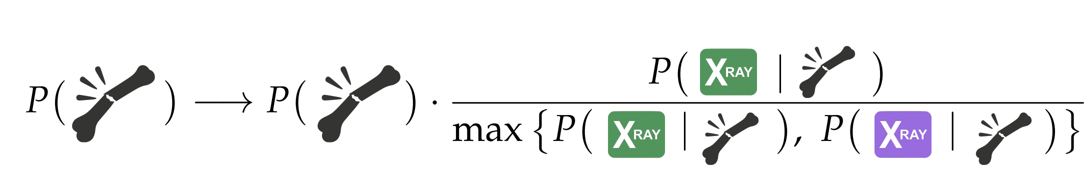
Main idea:

"discount" the weight of hypothesis by its likelihood ratio

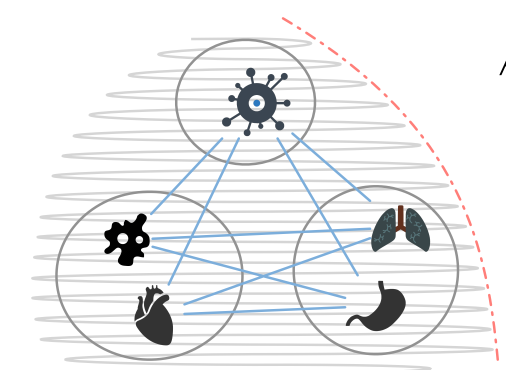


Likelihood ratio is small if the hypothesis is less likely to be "correct".

ECED adaptively pick tests that "discount" edge weight the most.



3 Account for Non-informative Tests



A test is non-informative if observing its outcome does not affect the distribution over the root-causes.

Introducing an "offset term" so that the gain of any non-informative tests is 0

ECED: Algorithmic Details

Initialization: $\mathcal{A} \leftarrow \emptyset;\ \mathbf{x}_\mathcal{A} \leftarrow \emptyset;$
$\forall \{\theta, \theta'\} \in \text{Edges, set } w_{\theta,\theta'} \leftarrow \Pr(\theta)\Pr(\theta');$

Loop: $e^* \in \arg\max_e \mathbb{E}_{x_e}[\textsc{Weight-Discounted}(x_e \mid \mathbf{x}_\mathcal{A}) - \textsc{Offset}(e \mid \mathbf{x}_\mathcal{A})];$
$\mathcal{A} \leftarrow \mathcal{A} \cup \{e^*\};\ \mathbf{x}_\mathcal{A} \leftarrow \mathbf{x}_\mathcal{A} \cup \{(e^*, x_{e^*})\};$

Main Results

Thm: ECED is competitive with the optimal policy that achieves some lower error probability.

It suffices to run ECED

$$\mathcal{O}\left(\frac{k}{(1-2\epsilon)^2} \cdot \left(\log \frac{kn}{\delta} \log \frac{n}{\delta}\right)^2\right)$$

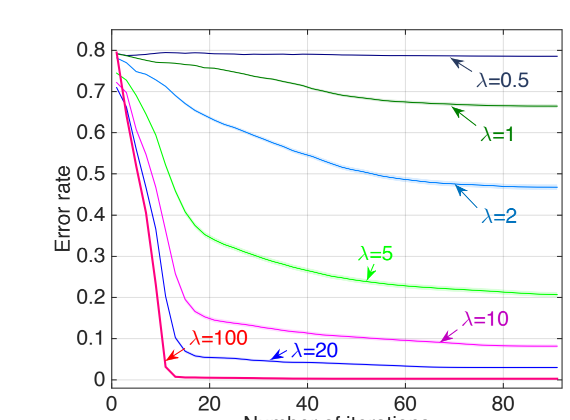
the worst-case cost of OPT which achieves expected error probability $\mathcal{O}\left(\frac{\delta}{(\log n \cdot \log(1/\delta))^2}\right)$

#root-causes

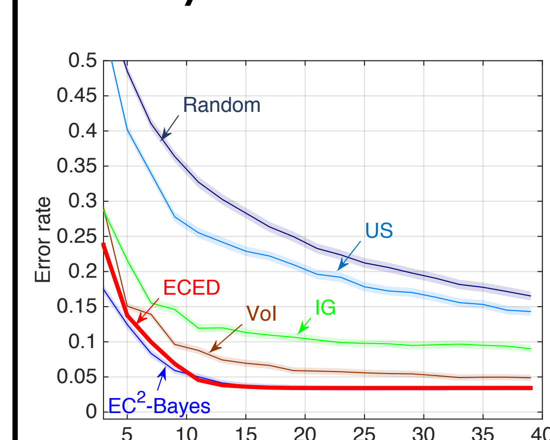
iterations to achieve expected error probability δ.

Experiments

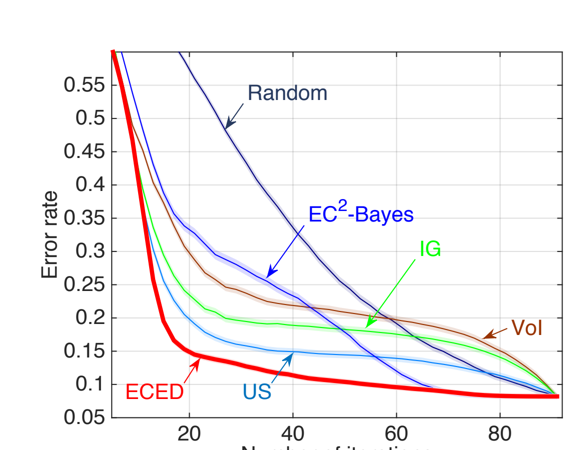
Random: Random Selection of Tests
US: Uncertainty Sampling
INF: Information Gain
VoI: Myopic Value of Information
EC²-Bayes: EC2 with Bayesian Updates
ECED: Equivalence Class Edge Discounting



Risky Choice Selection



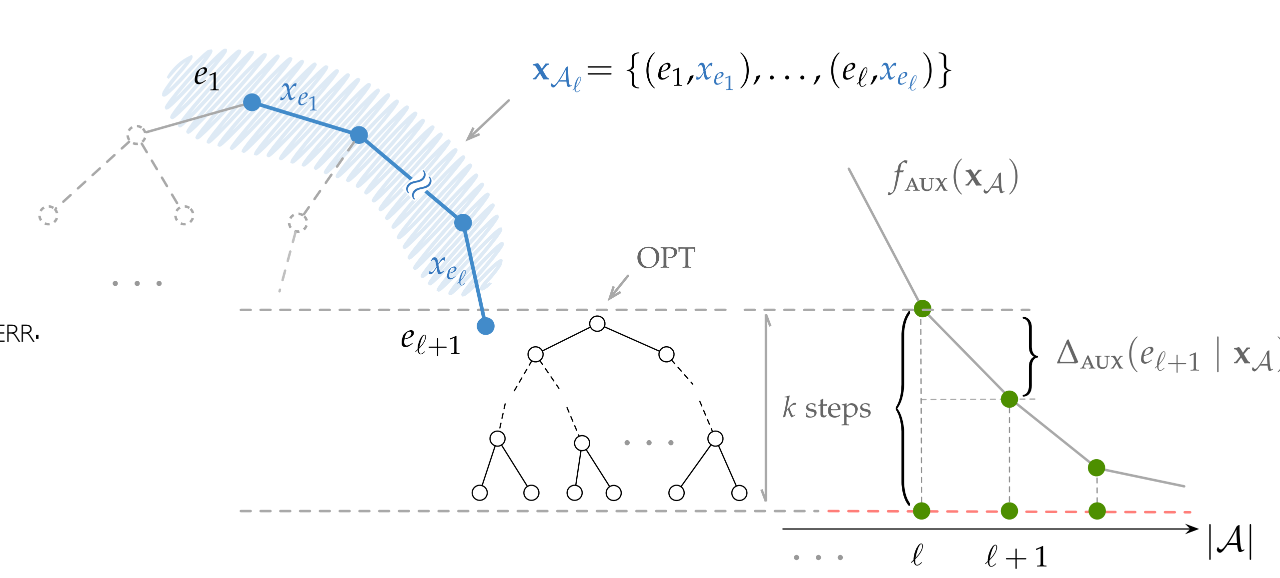
Comparison-based Search

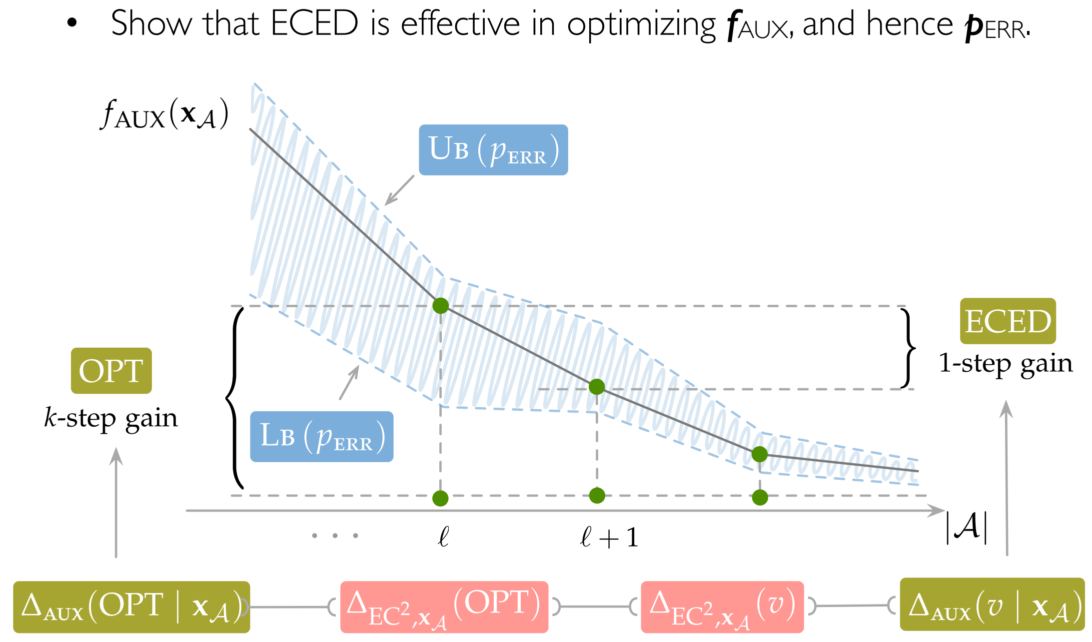


4 Key Lemma — Proof Sketch

Key idea

- Introduce an information-theoretic auxiliary function f_{AUX}
- Bound f_{AUX} against the target objective function p_{ERR}
- Show that ECED is effective in optimizing f_{AUX}, and hence p_{ERR}.





Our proof relates the 1-step gain of ECED (greedy) to the k-step gain of the optimal algorithm.

Lem: The 1-step gain of ECED regarding some auxiliary object function f_{AUX} is "significant", compared with the total gain of OPT.