Near-optimal Bayesian Active Learning with Correlated and Noisy Tests

Yuxin Chen, S. Hamed Hassani, Andreas Krause

1 Problem Setup

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

Example
Medical diagnosis

Root-cause
Θ = (θ₁, θ₂, ..., θₖ)

Target variable
Y = (y₁, y₂, ..., yₙ)

Tests T = {1, ..., m}, each with unit cost
Test outcomes X_{test} are correlated with Θ

(Noisy) Bayesian Active Learning

Def.
After performing a set of tests A, and observing outcomes X, the error probability of the MAP estimator of Y is

\[ p_{\text{err}}(X_A) = 1 - \max_{\pi \in \Pi} P(Y | X_A) \]

Given
Prior P(Θ)
Conditional probabilities P(X | Θ) ∈ [0, 1]
Mapping x from Θ to Y: π(Θ) = y

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

\[ \pi^* = \arg\min_{\pi \in \Pi} \mathbb{E}_{X_A}[\text{max}_{\theta \in \Theta} |\text{argmax}_{x \in [0,1]} P(x | \theta) - y|] \leq \delta \]

Observations seen by policy π, if the tests outcomes are realized as X

Existing Approaches

Greedy Heuristics
- Uncertainty sampling
- Maximize information gain [e.g., Linley’98]
- Maximize value of information [e.g., Howard’60]

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

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)

Extension to the Noisy Setting

[This paper — ECED: Equivalent Class Edge Discounting]

Main idea:
“discount” the weight of hypothesis by its likelihood ratio

\[ P(\theta) \rightarrow P(\theta) / \text{Likelihood ratio} \]

This test is not informative

\[ \max \{ P(X = x) \} \]

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

ECED adaptively picks tests that “discount” edges with the most.

Random: Random Selection of Tests
US: Uncertainty Sampling
IG: Information Gain
VC: Value of Information
EC²: EC² with Bayesian Updates
ECED: Equivalent Class Edge Discounting

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:
\[ A = \emptyset, \pi(t) = \pi_0 \]
\[ \gamma(t, t') \in \text{Edges, set } \gamma(t, t') \leftarrow P(\theta(t)) P(\theta(t')) \]

Loop:
\[ t \leftarrow \arg\max_{t \in [1, T]} \text{Weight-Discounted} \]
\[ = \text{Option}(\gamma(t), \pi(t)) \]
\[ A \leftarrow A \cup \{ \bigoplus_{(t', t)} \} ; \; \pi(t) \leftarrow \pi(t) \cup \{ \bigoplus(x, y) \} \]

Main Results

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

It suffices to run ECED the worst-case cost of OPT which achieves expected error probability

\[ \Theta \left( \frac{\log (\frac{1}{\delta})}{\log \frac{1}{\delta}} \right)^2 \]

iterations to achieve expected error probability δ.

Experiments

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 \( p_{\text{err}} \)
- Show that ECED is effective in optimizing \( f_{aux} \), and hence \( p_{\text{err}} \).

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.