Bayesian Adaptive Data Analysis: Challenges and Guarantees

Sam Elder
MIT Mathematics

Motivation

Adaptive data analysis studies the difficulties of adaptivity to discerning correct results from data analytic techniques. Formulated (DFHPRR ’14) as a game between a Curator with data and an Analyst studying a distribution by making adaptive queries:

What can adaptive queries do that static can’t? Interactive fingerprinting attack of HU’14/SU’14: Worst-case analyst could know the true distribution ̂p and quiz the curator about it.

What other problems arise in adaptive data analysis besides those resulting from an information asymmetry?

Bayesian Adaptive Data Analysis

Game between Analyst and Curator:
- Unknown distribution ̂p on universe X drawn from prior P.
- Curator receives P and n samples from ̂p.
- Analyst receives P and asks q statistical queries (averages of functions f_i : X → [0, 1]).
- Curator answers each query with answer a_i.
- Curator wins if all answers are approximately accurate on ̂p:
  \[ |a_i - E_{x\sim p} f_i(x)| < \epsilon \forall i. \]

Central question: How many samples n(\epsilon, \delta) does the curator need?

Lower Bound: New Problem

For a wide class of curator algorithms, there is a problem and adaptive analyst attack using O(n^3) queries which causes the curator to be 1/20-inaccurate on some query with 1/2 probability.

Posterior Uncertainty

Let C ⊂ F_2^n be a linear error-correcting code of size 2^k and distance d. Define model M_C as follows:
- Universe: [n] × F_2
- Population ̂p: For some codeword C ∈ C, uniform over (i, C_i).
- Prior: Uniform over all codewords.

This problem will make the curator uncertain:
- Posterior: only consistent hypotheses (2^k → 2^k−1 → ⋯ → 2 → 1)
- Error: ≥ d/2m on some query after ~ k samples.
- Justesen code has d ≈ m/10 and k ≈ m/4.

Slightly Correlated Queries

Curator might try to hide knowledge by:
- Adding noise to all answers (Laplacian/Gaussian).
- Rounding all answers (in a prior-sensitive way).
- Using a proxy distribution (PMW).

Augment problem with q − 1 uniformly randomly biased coins. Queries like this extract information about C_i:

\[ f_i : [n] \times F_2 \times ([0, 1])^{q-1} \rightarrow [0, 1] \]

for i = 0, ⋯, n − 1. Determine which subsets of the curator’s knowledge about C_i:

- C_i = 0
- C_i = ?
- C_i = 1

With O(m^2) queries, analyst learns C_i.

Upper Bound: Easy Scenario

If the prior P is a Dirichlet prior Dir(α_0, ⋯, α_k) for any α_0 > 0, then the posterior mean curator strategy achieves the static bound n = O(\frac{n}{\epsilon}).

Subgaussianity Condition

Recall: Against all static analysts, the empirical mean curator achieves

\[ n = \Theta\left( \frac{1}{\epsilon^2 \log \frac{q}{\delta}} \right). \]

In other words, q = δ \exp(Ω(nε^2)).

Proposition. If the curator’s posterior is O(1/n)-subgaussian with respect to any counting query, then the posterior mean curator achieves (1) against any adaptive analyst.

Beta Distribution Concentration

One family of priors: Dirichlet prior Dir(α_0, ⋯, α_k): β > n. For instance, α_0 = ⋯ = α_k = 1 is the uniform prior over the simplex.

- Conjugate family: After receiving n_i copies of i, posterior is Dir(α_i + n_i, ⋯, α_k + n_k).
- Converse adaptive bound (counting query v ∈ {0, 1}^k):

\[ \text{Dir}(α_i, ⋯, α_k) = \beta \left( \sum_{v=0}^1 \alpha_v, \sum_{v=1}^k \alpha_v \right) \]

Beta Distribution Concentration

The Beta(α, β) distribution is subgaussian with variance proxy

\[ \frac{1}{4(α + β) + 2} \]

Two very different proofs:
- Show E[λX] ≤ (nεX + λ^2 σ^2/2) by expanding in λ (i.e. bounding raw moments).
- Azuma’s inequality on the posterior mean evolution as n → ∞ (actually stronger).

Conclusions

I introduced the new BADA problem to understand what difficulties arise from utilizing adaptive data analytic techniques. Here are my answers:

- If posterior is uncertain, analyst can exploit.
- Obfuscation techniques can’t stop tricky queries.
- No need to obfuscate if posterior is subgaussian.
- If posterior is stable, curator can be confident.

Future Directions

There are several big open questions in this area:

- Are there general Bayesian algorithms that do better than those known previously (and in particular, beat my monster problem)?
- Can we profitably define further restrictions on the allowable analysts?
- Is there an even more general class of priors on which the posterior mean is accurate?
- What do positive results in the Bayesian setting correspond to in the original frequentist setting?
- To what extent do these results accurately describe data analysis in practice?
- Are other data techniques like cross-validation also vulnerable to the problems of adaptivity?

Acknowledgements

Thanks to Jon Kelner, Jerry Li, Adam Sealfon, Thomas Steinke, and the MIT learning theory group for numerous helpful conversations.

I received financial support from the United States Department of Defense through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program.

Comics from PhD Comics by Jorge Cham.

For More Information

- arXiv papers: 1604.02492 (lower bounds), 1611.00065 (upper bounds)
- E-mail: same@mit.edu