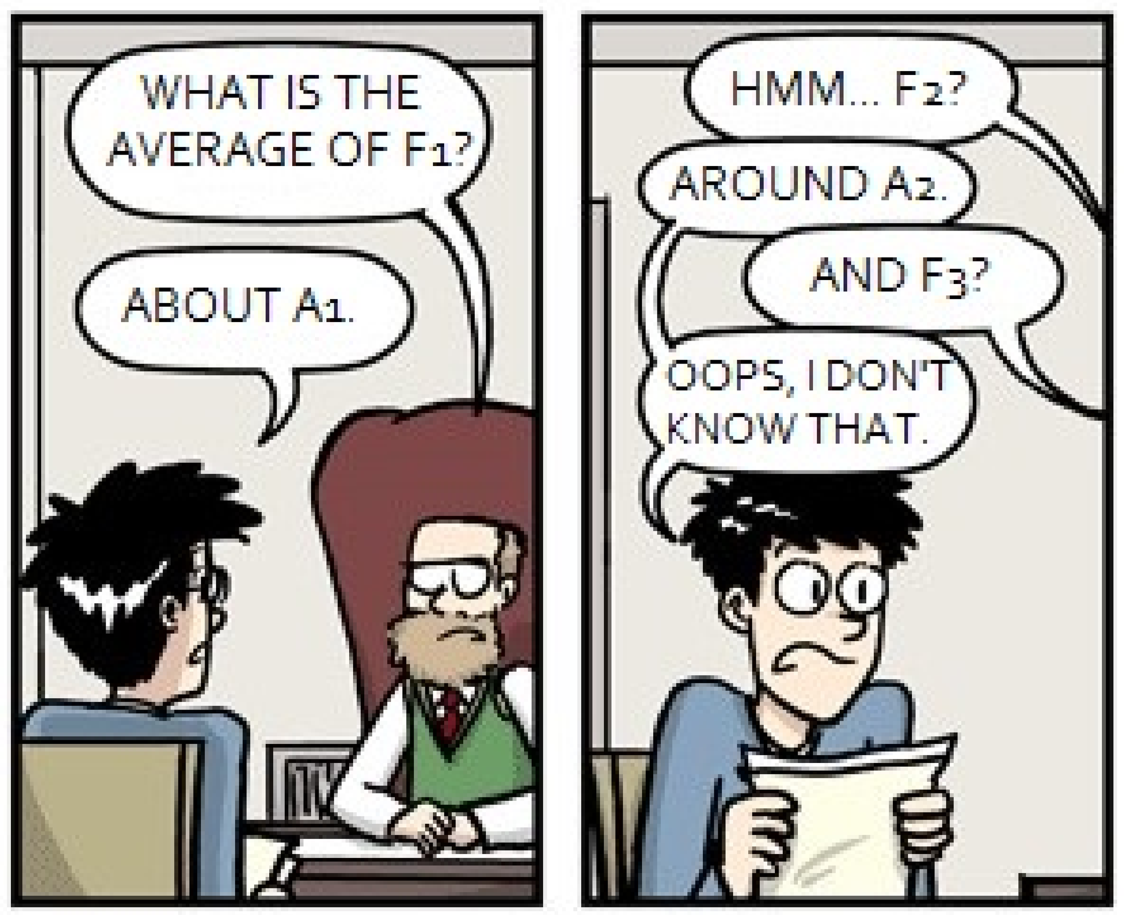
Bayesian Adaptive Data Analysis: Challenges and Guarantees

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 $\vec{p}$ 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 $\vec{p}$ on universe $\mathcal{X}$ **drawn from prior** $\mathcal{P}$.
- Curator receives $\mathcal{P}$ and n samples from $\vec{p}$.
- Analyst receives $\mathcal{P}$ and asks q statistical queries (averages of functions $f_i : \mathcal{X} \to [0,1]$).
- Curator answers each query with answer a_i.
- Curator wins if all answers are approximately accurate on $\vec{p}$:

$$\text{w.p. } 1-\delta, \quad |a_i - \mathbb{E}_{x\sim\vec{p}} f_i(x)| < \epsilon \; \forall i.$$

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

Lower Bound: New Problem

For a wide class of curator algorithms, there is a problem and adaptive analyst attack using $\tilde{O}(n^4)$ queries which causes the curator to be 1/20-inaccurate on some query with 1/2 probability.

Posterior Uncertainty

Let $\mathcal{C} \subset \mathbb{F}_2^m$ be a linear error-correcting code of size 2^k and distance d. Define model $\mathcal{M}_{\mathcal{C}}$ as follows:

- Universe: $[m] \times \mathbb{F}_2$
- Population $\vec{p}$: For some codeword $C \in \mathcal{C}$, uniform over (i, C_i).
- Prior: Uniform over all codewords.

This problem will make the curator uncertain:

- Posterior: only consistent hypotheses ($2^k \to 2^{k-1} \to \cdots \to \mathbf{2} \to 1$)
- Error $\geq d/2m$ on some query after $\sim k$ samples.
- Justesen code has $d \approx m/10$ and $k \approx m/4$.

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:

$$f_i : [m] \times \mathbb{F}_2 \times (\{0,1\})^{q-1} \to [0,1]$$

$$(y, z, x_1, \ldots, x_{q-1}) \mapsto \begin{cases} tz & \text{if } y = i \\ x_i & \text{if } y \neq i, \end{cases}$$

where $t = \Theta(1)$ matches one more success to the curator's knowledge about C_i:

$$\bullet\bullet\bullet\bullet\cdots\bullet\bullet \qquad C_i = 0$$
$$\bullet\bullet\bullet\cdots\bullet\bullet\bullet \qquad C_i = ?$$
$$\bullet\bullet\cdots\bullet\bullet\bullet\bullet \qquad C_i = 1$$

With $\tilde{O}(m^2)$ queries, analyst learns C_i.

Upper Bound: Easy Scenario

If the prior $\mathcal{P}$ is a Dirichlet prior $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$ for any $\alpha_i > 0$, then the posterior mean curator strategy achieves the static bound $n = O\left(\frac{1}{\epsilon^2}\log\frac{q}{\delta}\right)$.

Subgaussianity Condition

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

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

In other words, $q = \delta \exp(\Omega(n\epsilon^2))$.

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.
(Counting queries are averages of $f : \mathcal{X} \to \{0,1\}$.)

Beta Distribution Concentration

One family of priors: Dirichlet prior $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$, $\alpha_i > 0$. For instance, $\alpha_1 = \cdots = \alpha_k = 1$ is the uniform prior over the simplex.

- Conjugate family: After receiving n_i copies of i, posterior is $\mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$.
- Posterior after n samples is $\mathrm{Dir}(\alpha'_1, \ldots, \alpha'_k)$ with $\sum_i \alpha'_i > n$.
- With respect to counting query $v \in \{0,1\}^k$, $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$ is $\mathrm{Beta}\left(\sum_{v_i=0} \alpha_i, \sum_{v_i=1} \alpha_i\right)$.

Beta Distribution Concentration

The $\mathrm{Beta}(\alpha, \beta)$ distribution is subgaussian with variance proxy $\dfrac{1}{4(\alpha+\beta)+2}$.

Two very different proofs:

- Show $\mathbb{E}[\lambda X] \leq \exp(\lambda \mathbb{E}[X] + \lambda^2\sigma^2/2)$ by expanding in λ (i.e. bounding raw moments).
- Azuma's inequality on the posterior mean evolution as $n \to \infty$ (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