Bayesian Models and Information Symmetry in Adaptive Data Analysis

Adam Smith
Boston University
Simons Workshop on Adaptive Data Analysis
July 25, 2018
This talk

a “worst-case Bayesian” model of adaptive data analysis

• Importance of information symmetry
• Some lower bounds; more open problems
• Based on [Elder’16+] and discussions/work with Jon Ullman, Thomas Steinke, Kobbi Nissim, Uri Stemmer
Outline

• Adaptive linear query model

• Bayesian setting
  ➢ Definition
  ➢ The “only” problem: High-variance posteriors

• Game-theoretic perspective

• Lower bounds as estimation

• Lower bounds for the Bayesian model
Bayesian Setting

Worst-case model allows $A$ to choose $P$

- Known lower bounds rely on this!

What happens when we allow

- $M$ to see the code of $A$?
- $M$ to know “as much as” $A$ about $P$?

First attempt: what if $M$ knows $P$ exactly?

- Not interesting: $M_P$ can ignore data and answer $a_i = q_i(P)$

“Bayesian” setting:

- Consider a “hyperdistribution” $Gen$ that selects $P$
- What if $M$ and $A$ know $Gen$ but not $P$?

$$\inf_{Gen} \sup_{M} \sup_{P} \sup_{A} \mathbb{E}_{X \sim p^n \text{ coins}} \left( \max_{i} |a_i - q_i(P)| \right)$$

$$\sup_{Gen} \inf_{M} \sup_{A} \mathbb{E}_{X \sim p^n \text{ coins}} \left( \max_{i} |a_i - q_i(P)| \right)$$

$M$ has more power, so error can only go down
Bayesian Setting

- **Pros**
  - One model of “benign” analyst behavior
  - Captures widely-promoted statistical practice
    - c.f. *Interactive Data Analysis*, Bi, Markovic, Xia, Taylor, 2017
  - Maybe: algorithms with greater resistance to adaptive queries
    - Basically no nontrivial, universal lower bounds!

- **Cons**
  - May not model analyst with multiple data sets (composition)
  - Less robust?

- Nonadaptive queries
  \[ \frac{\sqrt{\log k}}{\sqrt{n}} \]

- Tracing queries
  \[ \frac{1}{\sqrt{n}} + \frac{\sqrt{k}}{n} \]

- Diff. Priv
  \[ \frac{4\sqrt{k}}{\sqrt{n}} \]

[Hardt, Ullman 14, Steinke, Ullman 15] [DFHPRR’15, BNSSSU’16]
“Bayesian” mechanisms

- Given $Gen$, and $X_1, ..., X_n \sim P \otimes n$:
  - Consider posterior distribution on $P|X$
    - Induces distribution on true mean $q(P)|X$

- Posterior-based mechanisms:
  On input $q_j$...
  - Posterior expected mean: $a_j = \mathbb{E}(q_j(P)|X)$
  - Noisy posterior mean: $a_j = \mathbb{E}(q_j(P)|X) + N(0, \sigma^2)$
  - Posterior confidence interval:
    $$a_j = \left( \text{quantile}_{0.05}(q_j(P)|X), \text{quantile}_{0.95}(q_j(P)|X) \right)$$

- Consistency [Elder]: When $P \sim Gen$ and $X \sim P \otimes n$, posterior-based mechanisms are “never wrong”
  - E.g. confidence interval captures $q_j(P)$ w.p. 90%
  - No matter if queries are adaptive, as long as queries depend on $P$ only via $X$.

Only possible problem: high-variance posterior

Example: biased coin flip

Posterior given 15 heads out of 25
Why do “tracing queries” fail?

• Set up
  ➢ Universe $U = \{1, \ldots, 2^{O(kn)}\}$
  ➢ $P$ is uniform over $T \subseteq U$, where $|T| = N$
  ➢ Mechanism sees $X \subseteq T$ of size $n$ but doesn’t know $T$

• Analyst knows $T$, chooses queries...
  ➢ At first: With bias $p_j$ on $T$, but bias $1/2$ on $U \setminus T$
    • Key fact: Accurate answers based only on $X$ leak information about $X$
    • Large universe makes it hard to identify $T$
    • Analysts learns $\hat{X} \subseteq X$
  ➢ Later: with bias $p_j$ on $T \setminus \hat{X}$, but bias $1/2$ on $\hat{X} \cup (U \setminus T)$

• Bayesian setting
  ➢ Mechanism knows $T$, can ignore $X$
Impossibility Results

Only possible problem: high-variance posterior

What can we say about variance?

• Nonadaptive linear queries
  ➢ Posterior mean/median have error $O(\log k / \sqrt{n})$

• How many queries can we answer adaptively?
  ➢ Empirical mean + Gaussian: can answer $\Omega(n^2)$
  ➢ Posterior mean: __________ $O(n)$ queries cause problems
  ➢ Posterior mean + Gaussian: _ $O(n^{2.5})$ queries [S, Steinke, Ullman]
  ➢ Posterior mean + arbitrary: _ $O(n^4)$ queries [Elder]
  ➢ Poly-time mechanisms: ______ $O(n^2)$ queries [Nissim, Stemmer]
  ➢ General mechanisms: $2^{O(n)}$ queries—same as for nonadaptive 😞
Outline

• Adaptive linear query model

• Bayesian setting
  ➢ Definition
  ➢ The “only” problem: High-variance posteriors

• Game-theoretic perspective

• Lower bounds as estimation

• Lower bounds for the Bayesian model
Three player game

1. Population player generates $P$
   - Random strategy is “hyperdistribution” over $P$

2. Mechanism player selects (randomized) $M$

3. Analyst selects (randomized) $A$

\[ Value = \mathbb{E}_{\text{everything}} \left( \max_i |a_i - q_i(P)| \right) \]

- “Worst-case” distribution model [DFHPRR/HU]:
  - First randomized $M$, then $(P, A)$ together
  \[ \inf_M \sup_P \sup_A \mathbb{E}_{X \sim p^n} \left( \max_i |a_i - q_i(P)| \right) \]
  - This is a Nash equilibrium, so can switch order: first joint distribution over $(P, A)$, then $M$
Three player game

1. Population player generates $P$
   ➢ Random strategy is “hyperdistribution” over $P$

2. Mechanism player selects (randomized) $M$

3. Analyst selects (randomized) $A$

\[ Value = \mathbb{E}_{every} \left( \max_i |a_i - q_i(P)| \right) \]

• Bayesian model [Elder]
  ➢ First $Gen$, then $M$ and $A$ separately.
    • $P$ and $A$ selected independently
  ➢ For each $Gen$, Nash equilibrium allows swapping $M, A$
• How do the values of these games compare?
  ➢ Bayesian setting is easier for mechanism
  ➢ So
    \[
    value(\text{Bayesian}) \leq value(\text{worst-case})
    \]

• Bayesian setting: May as well show code of analyst to mechanism
Outline

• Adaptive linear query model

• Bayesian setting
  ➢ Definition
  ➢ The “only” problem: High-variance posteriors

• Game-theoretic perspective

• Lower bounds as estimation

• Lower bounds for the Bayesian model
Proving lower bounds corresponds to finding $Gen$, $f$ and $\alpha$.

- Positive result: $k$ adaptive queries to SQ oracle allow approximating $f(P)$.
- Negative result: $n$ samples from $P$ do not.

Current lower bounds involve extra side information visible to $A$ but not oracle.
Outline

• Adaptive linear query model

• Bayesian setting
  ➢ Definition
  ➢ The “only” problem: High-variance posteriors

• Game-theoretic perspective

• Lower bounds as estimation

• Lower bounds for the Bayesian model
What if analyst sees the raw data?

Example 1: Coin flips
• Domain = \{0,1\}^d
  ➢ Coordinates are independent
  ➢ \(P\) described by biases \(p_1, ..., p_d\)
  ➢ Gen: Each bias \(p_j \in R \{\frac{1}{3}, \frac{2}{3}\}\), i.i.d.
• If some coordinate has \(n/2\) ones, then posterior distribution is \(\{\frac{1}{3}, \frac{2}{3}\}\)
  ➢ Analyst finds a bad query (w.h.p.) when \(d = 2^{\Omega(n)}\)

Example 2: Parities
• Domain = \{0,1\}^d
  ➢ \(P_z\) : Uniform on \(\{u: z \odot u = 0\}\)
  ➢ Gen: select \(Z \in R \{0,1\}^d\)
• If \(x\) has \(d - 1\) linearly independent vectors,
  ➢ then \(Z|x\) is uniform \(\{z_1, z_2\}\)
  ➢ Analyst can ask query with different values on \(z_1, z_2\)
• If \(n = d\), probability of exactly \(d - 1\) linear constraints is \(1/4\)
What about using linear queries?

- Replace parities with coding construction [Elder]
- Set up
  - Consider linear error-correcting code $C \subseteq F_2^N$, dimension $d$
  - $U = [N] \times F_2$
  - $Gen$: Select $c \in R C$, output $P_c$ uniform on \{(i, c_i) : i \in [N]\}
- When can we find high-variance queries?
  - $X$ gives a set of linear constraints on $c$
  - Suppose they have rank $d - 1$
    - Then $c|x$ is uniform on $\{c_1, c_2\}$ $\Rightarrow$ bad query
    - $\Pr(rank(x) = d - \Omega(1)) = \Theta(1/\sqrt{n})$
- How can we extract $x$ from answers to linear queries?
  - Let $sh(x) \in \{0, -1, +1\}^N$ denote “signed histogram” for $x$
    - $sh(x)_i = 0$ if position is absent, and $\pm 1$ otherwise
  - Posterior distribution $sh(P)|x$ equals $\frac{1}{N}sh(x)$
  - Ask linear queries on $sh$.

Posterior mean + arbitrary: $\bar{O}(n^4)$ queries
Posterior mean + Gaussian: $\bar{O}(n^{2.5})$ queries
Computationally bounded mechanisms

• Suppose $M$ is polynomial time
• Use public-key crypto to conceal $T$ in tracing attack [Nissim Stemmer]
  ➢ Public info: $pk_1, pk_2, \ldots, pk_n$
  ➢ $U = \{(i, sk_i): i = 1, \ldots, N\}$
  ➢ $X = \{(i, sk_i): i \in S\}$ where $|S| = n$
  ➢ Attacker encrypts query values with public keys
    • Mechanism sees only query restricted to $X$

• Theorem: In Bayesian setting, polynomial-time mechanisms can answer $k = \tilde{O}(n^2)$ in worst case
**Impossibility Results**

**Only** possible problem: high-variance posterior

What can we say about variance?

• Nonadaptive linear queries
  - Posterior mean/median have error $O\left(\sqrt{\log k / n}\right)$

• How many queries can we answer adaptively?
  - Empirical mean + Gaussian: can answer $\Omega(n^2)$
  - Posterior mean: _______ $O(n)$ queries cause problems
  - Posterior mean + Gaussian: _ $O(n^{2.5})$ queries [S,Steinke,Ullman]
  - Posterior mean + arbitrary: _ $O(n^4)$ queries [Elder]
  - Poly-time mechanisms: ______ $O(n^2)$ queries [Nissim,Stemmer]
  - General mechanisms: $2^{O(n)}$ queries—same as for nonadaptive 😞
Outline

• Adaptive linear query model

• Bayesian setting
  ➢ Definition
  ➢ The “only” problem: High-variance posteriors

• Game-theoretic perspective

• Lower bounds as estimation

• Lower bounds for the Bayesian model
### Bayesian setting

#### Pros
- One model of “benign” analyst behavior
- Captures widely-promoted statistical practice
  - Interactive Data Analysis, Bi, Markovic, Xia, Taylor
- Maybe: algorithms with greater resistance to adaptive queries
  - Basically no nontrivial, universal lower bounds!

#### Cons
- May not model analyst with multiple data sets (composition)
- Less robust?

#### Open: A better understanding of the setting