Bayesian Adaptive Data Analysis: Difficulties and Guarantees

Sam Elder (MIT)

December 9, 2016

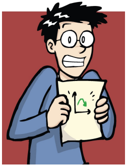
Central question: What makes adaptivity difficult?

Central question: What makes adaptivity difficult?

Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:

Central question: What makes adaptivity difficult?

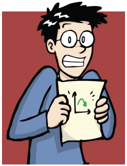
Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:



Curator

Central question: What makes adaptivity difficult?

Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:

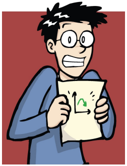


Curator

Analyst

Central question: What makes adaptivity difficult?

Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:



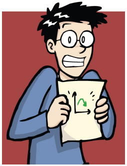
Curator

Receives n samples from $\vec{p}$.

Analyst

Central question: What makes adaptivity difficult?

Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:



Curator

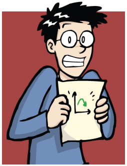
Receives n samples from $\vec{p}$.

Analyst

Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0, 1]$).

Central question: What makes adaptivity difficult?

Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:



Curator

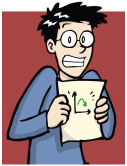
Receives n samples from $\vec{p}$.
Answers w/estimates a_i of each query.

Analyst

Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0,1]$).

Central question: What makes adaptivity difficult?

Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:



Curator

Receives n samples from $\vec{p}$.
Answers w/estimates a_i of each query.

Analyst

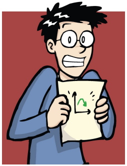
Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0,1]$).

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: What makes adaptivity difficult?

Original game formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$. Two players:



Receives n samples from $\vec{p}$.
Answers w/estimates a_i of each query.

Curator

Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0,1]$).

Analyst

Curator wins if all answers are approximately accurate on $\vec{p}$:

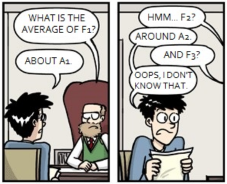
$$\text{w.p. } 1-\delta, \quad |a_i - \mathbb{E}_{x\sim\vec{p}} f_i(x)| < \epsilon \; \forall i.$$

In terms of ϵ, δ, q, how many samples n does the curator need?

Central question: What makes adaptivity difficult?

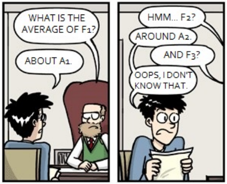
Static queries: $n = \Theta\left(\frac{1}{\epsilon^2} \log \frac{q}{\delta}\right)$.

Central question: What makes adaptivity difficult?



Static queries: $n = \Theta\left(\frac{1}{\epsilon^2} \log \frac{q}{\delta}\right)$.
What can an adaptive analyst do?

Central question: What makes adaptivity difficult?

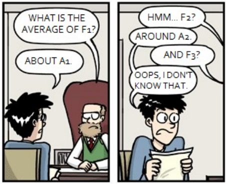


Static queries: $n = \Theta\left(\frac{1}{\epsilon^2}\log\frac{q}{\delta}\right)$.
What can an adaptive analyst do?

Interactive fingerprinting attack [HU'14,SU'14]

With $q = O_{\epsilon,\delta}(n^2)$ queries, find what data curator knows and ask about unseen data.

Central question: What makes adaptivity difficult?



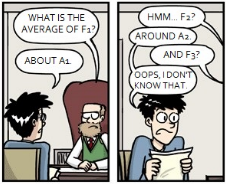
Static queries: $n = \Theta\left(\frac{1}{\epsilon^2} \log \frac{q}{\delta}\right)$.
What can an adaptive analyst do?

Interactive fingerprinting attack [HU'14,SU'14]

With $q = O_{\epsilon,\delta}(n^2)$ queries, find what data curator knows and ask about unseen data.

Problem: This requires the analyst to know $\vec{p}$.

Central question: What makes adaptivity difficult?



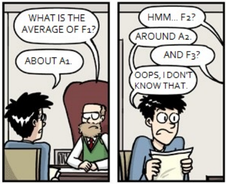
Static queries: $n = \Theta\left(\frac{1}{\epsilon^2} \log \frac{q}{\delta}\right)$.
What can an adaptive analyst do?

Interactive fingerprinting attack [HU'14,SU'14]

With $q = O_{\epsilon,\delta}(n^2)$ queries, find what data curator knows and ask about unseen data.

Problem: This requires the analyst to know $\vec{p}$. The analyst is asking queries he already knows the correct answers to!

Central question: What makes adaptivity difficult?



Static queries: $n = \Theta\left(\frac{1}{\epsilon^2}\log\frac{q}{\delta}\right)$.
What can an adaptive analyst do?

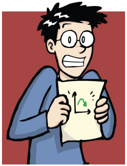
Interactive fingerprinting attack [HU'14,SU'14]

With $q = O_{\epsilon,\delta}(n^2)$ queries, find what data curator knows and ask about unseen data.

Problem: This requires the analyst to know $\vec{p}$. The analyst is asking queries he already knows the correct answers to!

Story: Adaptivity is hard because the analyst might already know the answers and quiz the curator.

Original formulation (DFHPRR '14):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$.



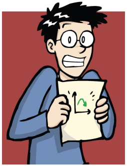
Curator

Receives n samples from $\vec{p}$.
Answers w/estimates a_i of each query.

Analyst

Receives $\vec{p}$.
Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0, 1]$).

Bayesian formulation (E '16):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$, public prior $\mathcal{P}$ over $\vec{p}$.



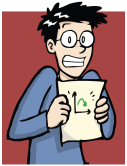
Curator

Receives $\mathcal{P}$ and n samples from $\vec{p}$.
Answers w/estimates a_i of each query.

Analyst

Receives $\mathcal{P}$.
Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0, 1]$).

Bayesian formulation (E '16):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$, public prior $\mathcal{P}$ over $\vec{p}$.



Curator

Receives $\mathcal{P}$ and n samples from $\vec{p}$. Answers w/estimates a_i of each query.

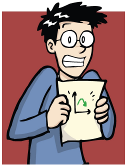
Analyst

Receives $\mathcal{P}$. Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0, 1]$).

Properties:

- Mandates information symmetry.

Bayesian formulation (E '16):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$, public prior $\mathcal{P}$ over $\vec{p}$.



Curator

Receives $\mathcal{P}$ and n samples from $\vec{p}$. Answers w/estimates a_i of each query.

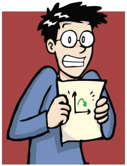
Analyst

Receives $\mathcal{P}$. Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0, 1]$).

Properties:

- Mandates information symmetry.
- Perhaps reasonable model of scientific research in practice.

Bayesian formulation (E '16):
Unknown distribution $\vec{p}$ on universe $\mathcal{X}$, public prior $\mathcal{P}$ over $\vec{p}$.



Receives $\mathcal{P}$ and n samples from $\vec{p}$. Answers w/estimates a_i of each query.

Curator

Receives $\mathcal{P}$.
Asks q *statistical queries* (averages of $f_i : \mathcal{X} \to [0, 1]$).

Analyst

Properties:

- Mandates information symmetry.
- Perhaps reasonable model of scientific research in practice.
- Prior as analysis tool.

Main negative result (new problem):

Main negative result (new problem):

Theorem

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.

Main negative result (new problem):

Theorem

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.

Main positive result (worry-free contexts):

Main negative result (new problem):

Theorem

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.

Main positive result (worry-free contexts):

Theorem

If the posterior is $O(1/n)$-subgaussian with respect to any query (e.g. if $\mathcal{P}$ is a Dirichlet prior), then the posterior mean curator strategy achieves the static bound $n = O\left(\frac{1}{\epsilon^2}\log\frac{q}{\delta}\right)$.

Are there queries a Bayesian curator can't estimate?

Are there queries a Bayesian curator can't estimate?

Definition

Let $\mathcal{C} \subset \mathbb{F}_2^m$ be a linear error-correcting code of size 2^k with distance d. Model $\mathcal{M}_\mathcal{C}$ is defined 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.

Are there queries a Bayesian curator can't estimate?

Definition

Let $\mathcal{C} \subset \mathbb{F}_2^m$ be a linear error-correcting code of size 2^k with distance d. Model $\mathcal{M}_{\mathcal{C}}$ is defined 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.

Properties:

- Posterior: only consistent hypotheses $(2^k \to \cdots \to \mathbf{2} \to 1)$

Are there queries a Bayesian curator can't estimate?

Definition

Let $\mathcal{C} \subset \mathbb{F}_2^m$ be a linear error-correcting code of size 2^k with distance d. Model $\mathcal{M}_\mathcal{C}$ is defined 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.

Properties:

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

Are there queries a Bayesian curator can't estimate?

Definition

Let $\mathcal{C} \subset \mathbb{F}_2^m$ be a linear error-correcting code of size 2^k with distance d. Model $\mathcal{M}_\mathcal{C}$ is defined 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.

Properties:

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

Curator can be $1/20$-uncertain after arbitrarily many samples.

Can the curator hide his uncertainty from the analyst?

Can the curator hide his uncertainty from the analyst?

Obfuscation techniques:

- Add noise to all answers (Laplacian/Gaussian).
- Round all answers (in a prior-sensitive way).
- Use a proxy distribution (PMW).

Can the curator hide his uncertainty from the analyst?

Obfuscation techniques:

- Add noise to all answers (Laplacian/Gaussian).
- Round all answers (in a prior-sensitive way).
- Use a proxy distribution (PMW).

We introduce a new attack which extracts information from all of these techniques using only $\tilde{O}(n^4)$ queries.

Can the curator hide his uncertainty from the analyst?

Obfuscation techniques:

- Add noise to all answers (Laplacian/Gaussian).
- Round all answers (in a prior-sensitive way).
- Use a proxy distribution (PMW).

We introduce a new attack which extracts information from all of these techniques using only $\tilde{O}(n^4)$ queries.

- Error-correcting code problem (previous slide).

Can the curator hide his uncertainty from the analyst?

Obfuscation techniques:

- Add noise to all answers (Laplacian/Gaussian).
- Round all answers (in a prior-sensitive way).
- Use a proxy distribution (PMW).

We introduce a new attack which extracts information from all of these techniques using only $\tilde{O}(n^4)$ queries.

- Error-correcting code problem (previous slide).
- Add $q-1$ uniformly randomly biased coins.

Can the curator hide his uncertainty from the analyst?

Obfuscation techniques:

- Add noise to all answers (Laplacian/Gaussian).
- Round all answers (in a prior-sensitive way).
- Use a proxy distribution (PMW).

We introduce a new attack which extracts information from all of these techniques using only $\tilde{O}(n^4)$ queries.

- Error-correcting code problem (previous slide).
- Add $q-1$ uniformly randomly biased coins.
- Ask a series of *slightly correlated queries* like these:

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

In what situations is adaptivity not a concern at all?

In what situations is adaptivity not a concern at all?

Recall: Against all static analysts, empirical mean curator achieves

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

In what situations is adaptivity not a concern at all?

Recall: Against all static analysts, empirical mean curator achieves

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

Proposition

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

In what situations is adaptivity not a concern at all?

Recall: Against all static analysts, empirical mean curator achieves

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

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 functions $f : \mathcal{X} \to \{0, 1\}$.)

In what situations is adaptivity not a concern at all?

In what situations is adaptivity not a concern at all?

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

In what situations is adaptivity not a concern at all?

One family of priors: Dirichlet prior $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$, $\alpha_i > 0$.
(e.g. $\alpha_1 = \cdots = \alpha_k = 1$ 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 Beta $\left(\sum_{v_i=0} \alpha_i, \sum_{v_i=1} \alpha_i\right)$.

In what situations is adaptivity not a concern at all?

One family of priors: Dirichlet prior $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$, $\alpha_i > 0$.
(e.g. $\alpha_1 = \cdots = \alpha_k = 1$ 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 $\mathrm{Beta}\left(\sum_{v_i=0} \alpha_i, \sum_{v_i=1} \alpha_i\right)$.

Theorem

The $\mathrm{Beta}(\alpha, \beta)$ *distribution is* $\frac{1}{4(\alpha+\beta)+2}$*-subgaussian.*

What makes adaptivity problematic?

What makes adaptivity problematic?

- Old answer: Standard statistical guarantees don't apply.

What makes adaptivity problematic?

- Old answer: Standard statistical guarantees don't apply.
- New [HU'14, SU'14] answer: Analysts who quiz the curator about their own extra knowledge.

What makes adaptivity problematic?

- Old answer: Standard statistical guarantees don't apply.
- New [HU'14, SU'14] answer: Analysts who quiz the curator about their own extra knowledge.
- Newer: Posterior uncertainty, possible information leakage.

What makes adaptivity problematic?

- Old answer: Standard statistical guarantees don't apply.
- New [HU'14, SU'14] answer: Analysts who quiz the curator about their own extra knowledge.
- Newer: Posterior uncertainty, possible information leakage.
- Newest: Posterior mean instability, over-leveraged data.

What makes adaptivity problematic?

- Old answer: Standard statistical guarantees don't apply.
- New [HU'14, SU'14] answer: Analysts who quiz the curator about their own extra knowledge.
- Newer: Posterior uncertainty, possible information leakage.
- Newest: Posterior mean instability, over-leveraged data.

Thank you.

ArXiv: 1604.02492 (lower bounds), 1611.00065 (upper bounds)
All comics from *PhD Comics* by Jorge Cham.