Bayesian Models and Information Symmetry in Adaptive Data Analysis

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

This talk

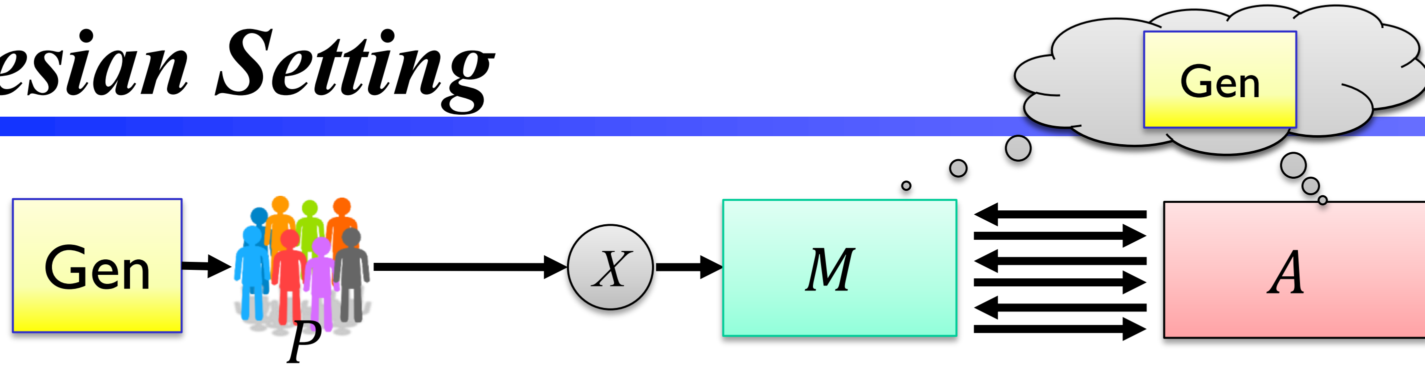
a "worst-case Bayesian" model of adaptive data analysis

- 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? later
- 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:

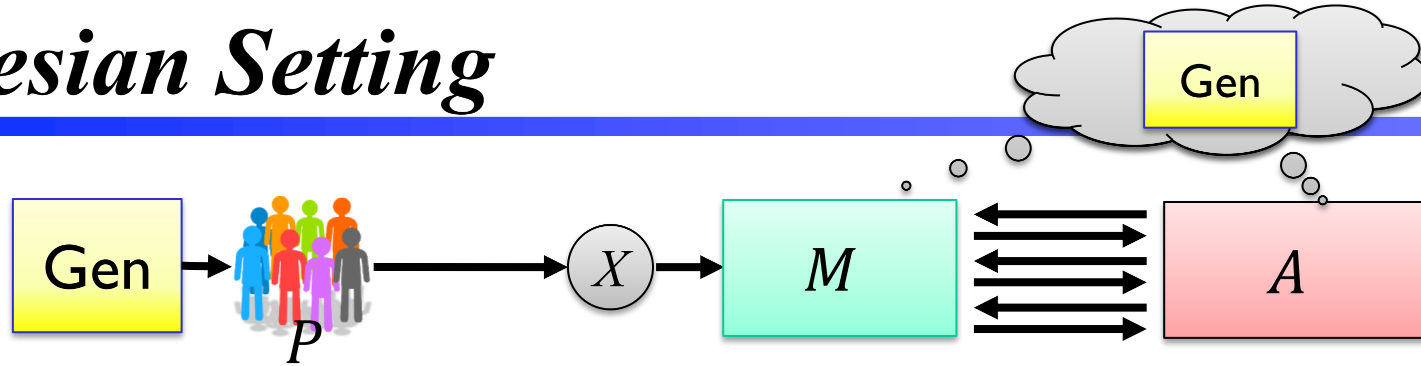
- Consider a "hyperdistribution" Gen that selects P
- What if M and A know Gen but not P?

$$\cancel{\inf_{M} \sup_{P} \sup_{A} \mathbb{E}_{\substack{X\sim P^n \\ coins}} \left(\max_i |a_i - q_i(P)| \right)}$$

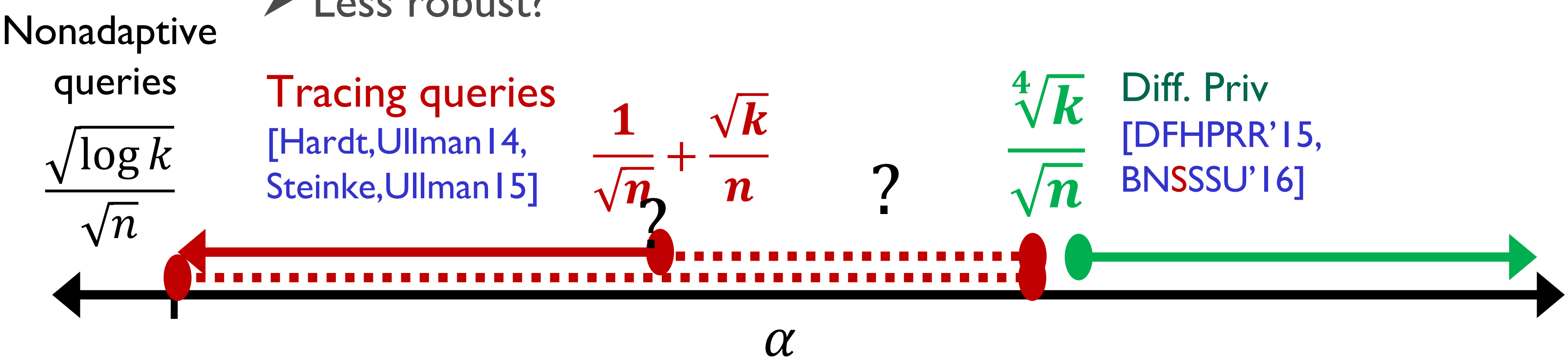
$$\sup_{Gen} \inf_{M} \sup_{A} \mathbb{E}_{\substack{X\sim P^n \\ coins}} \left(\max_i |a_i - q_i(P)| \right)$$

M has more power, so error can only go down

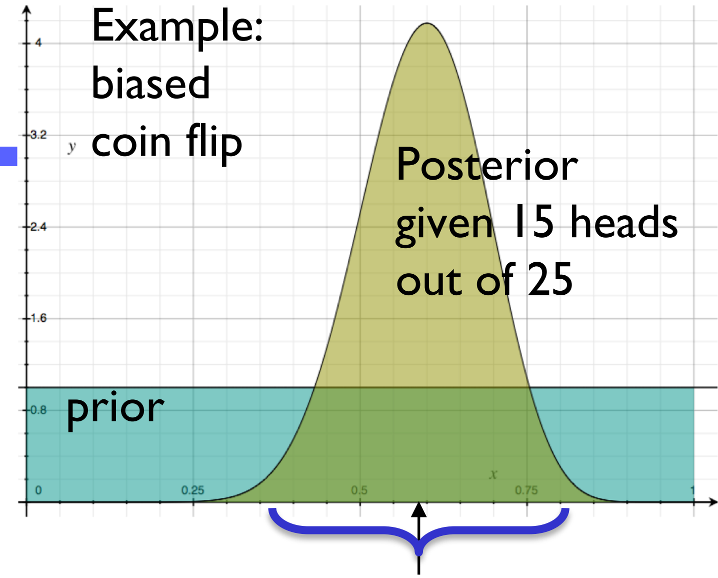
Bayesian Setting



- Pros
  - One model of "benign" analyst behavior
  - Captures widely-promoted statistical practice
    - c.f. *Inferactive 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?



“Bayesian” mechanisms



- Given *Gen*, and $X_1, \ldots, X_n \sim P^{\otimes n}$:
  - Consider posterior distribution on $P|X$
    - Induces distribution on true mean $q(P)|X$
- Posterior-based mechanisms:
  On input $q_j \ldots$
  - Posterior expected mean: $a_j = \mathbb{E}(q_j(P)|X)$
  - Noisy posterior mean: $a_j = \mathbb{E}\big(q_j(P)\big|X\big) + \mathrm{N}(0, \sigma^2)$
  - Posterior confidence interval:
    $$a_j = \Big(quantile_{0.05}\big(q_j(P)\big|X\big), quantile_{0.95}\big(q_j(P)\big|X\big)\Big)$$
- **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.

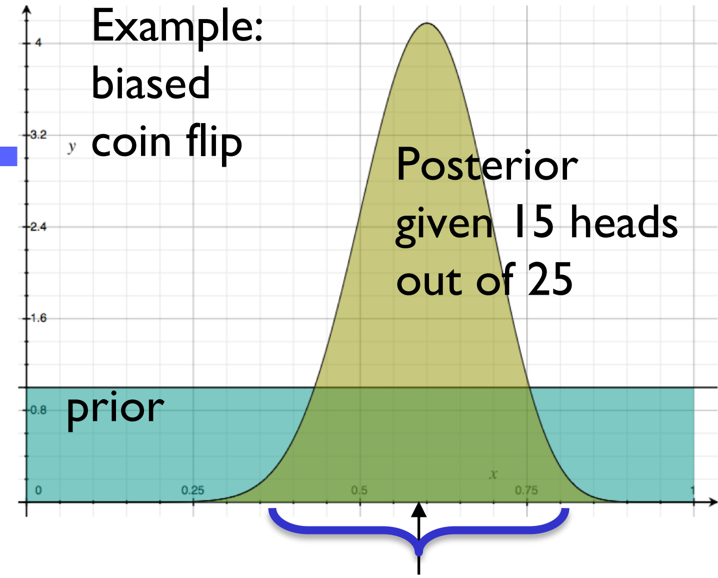
Only possible problem: high-variance posterior

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 $U \backslash \mathrm{T}$
    - 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 $T \backslash \hat{\mathrm{X}}$, but bias $1/2$ on $\hat{X} \cup (U \backslash \mathrm{T})$
- Bayesian setting
  - Mechanism knows T, can ignore X

Impossibility Results

Only possible problem: high-variance posterior



What can we say about variance?

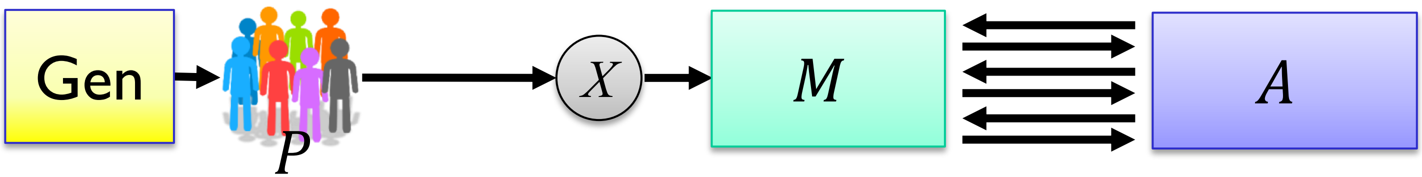
- Nonadaptive linear queries
  - ➢ Posterior mean/median have error $O(\log k/\sqrt{n})$
- How many queries can we answer adaptively?
  - ➢ Empirical mean + Gaussian: can answer $\Omega(n^2)$
  - ➢ Posterior mean: ____________ $O(n)$ queries cause problems
  - ➢ Posterior mean + Gaussian: _ $O(n^{2.5})$queries [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



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

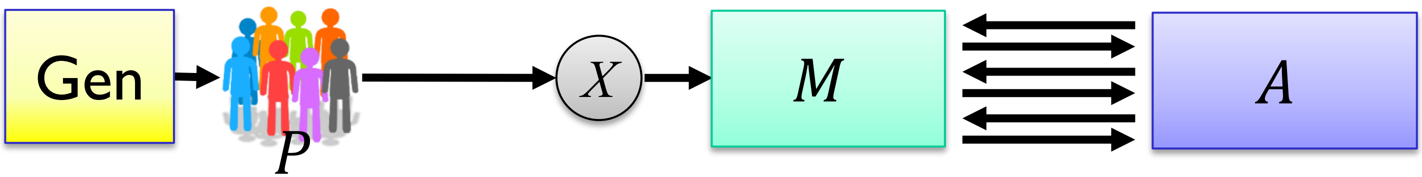
- "Worst-case" distribution model [DFHPRR/HU]:
  - First randomized M, then (P, A) together

$$\inf_M \sup_P \sup_A \; \mathbb{E}_{\substack{X\sim P^n \\ coins}} \left(\max_i |a_i - q_i(P)| \right)$$

  - 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



$$Value = \mathop{\mathbb{E}}_{everything}\left(\max_i |a_i - q_i(P)| \right)$$

- 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

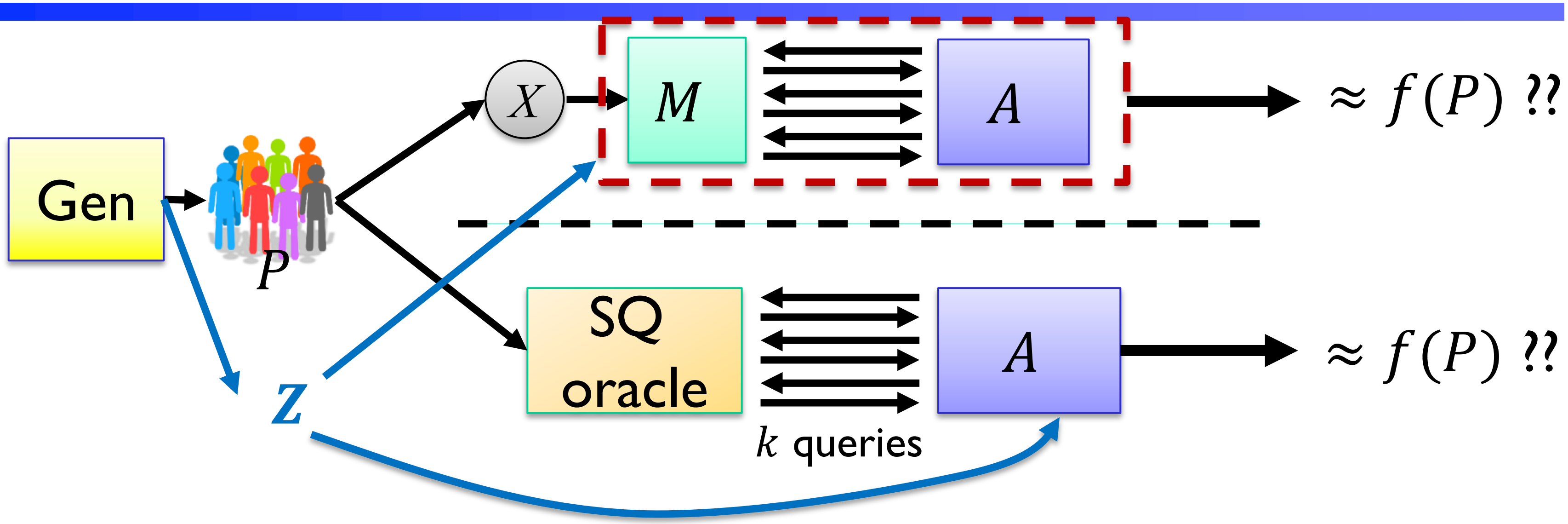
$$value(Bayesian) \leq value(worst - case)$$

- 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

Lower Bounds as Estimation

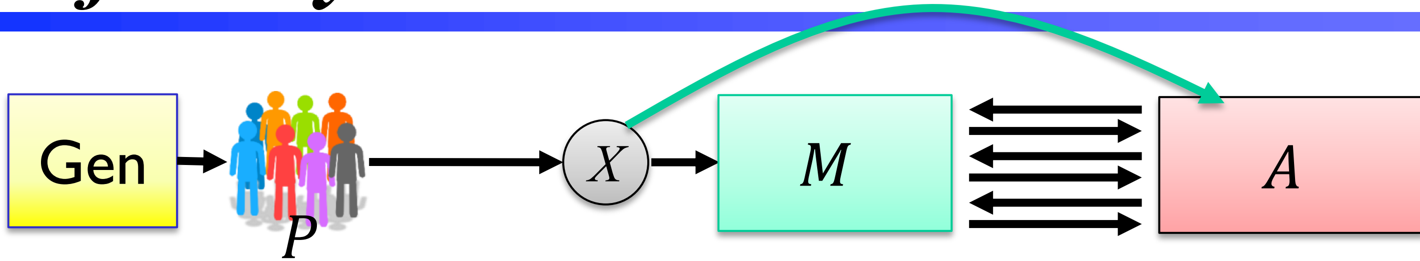


- Proving lower bounds corresponds to finding Gen, f and
  - Positive result: k adaptive queries to SQ oracle allow approximating $f(P)$
  - Negative result: n samples from P do not.
- Current lower bounds involve extra side information visible to A but not oracle

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 $p_1, \ldots, p_d$
  - *Gen*: Each bias $p_j \in_R \left\{\frac{1}{3}, \frac{2}{3}\right\}$, i.i.d.
- If some coordinate has $n/2$ ones, then posterior distribution is $\left\{\frac{1}{3}, \frac{2}{3}\right\}$
  - Analyst finds a bad query (w.h.p.) when $d = 2^{\Omega(n)}$

Don't know how to find a "bad" coordinate using linear queries

Example 2: Parities

- Domain = $\{0,1\}^d$
  - P_z : Uniform on $\{u\colon z \odot u = 0\}$
  - 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 $1/4$

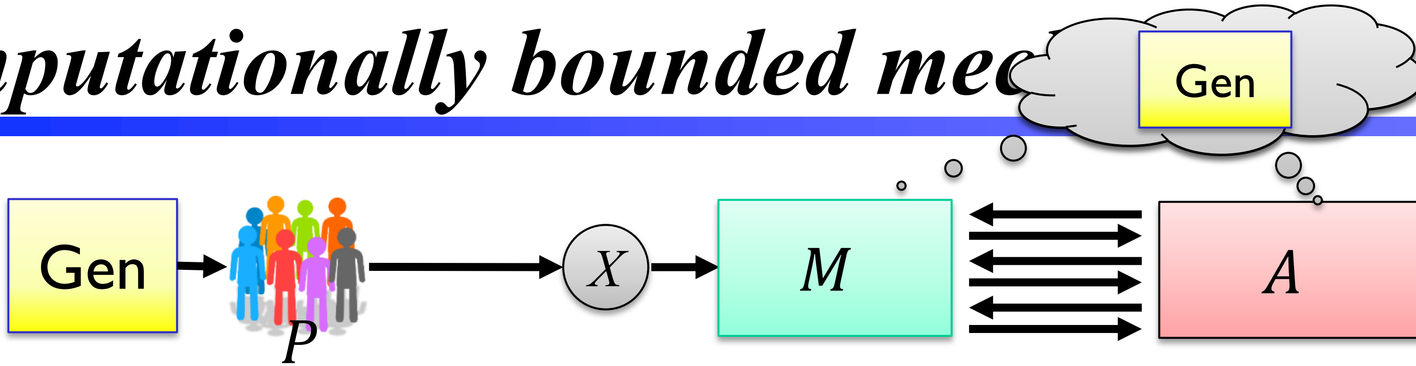
Can't extract info about z using poly many SQ queries

What about using linear queries?

- Replace parities with coding construction [Elder]
- Set up
  - Consider linear error-correcting code $C \subset F_2^N$, 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 $\{c_1, c_2\}$ $\implies$ bad query
  - $\Pr\left(rank(x) = d - \Omega(1)\right) = \Theta(1/\sqrt{n})$
- 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 ± 1 otherwise
  - Posterior distribution $sh(P)|x$ equals $\frac{1}{N} sh(x)$
  - Ask linear queries on sh.

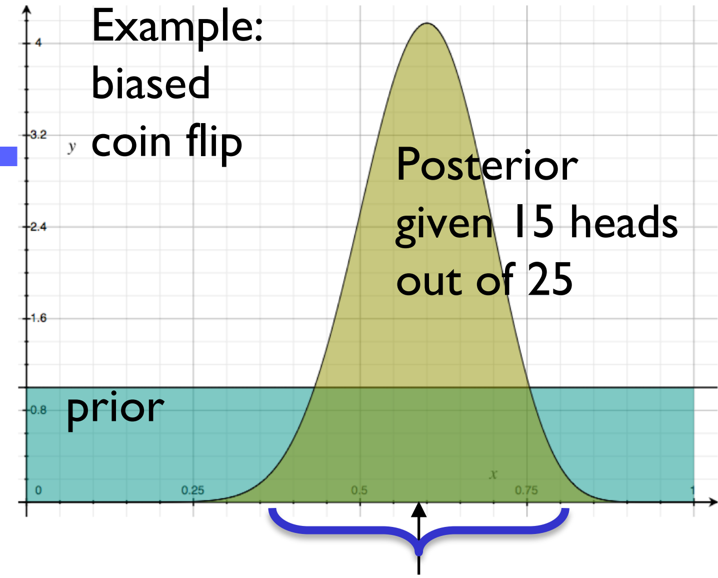
Posterior mean + arbitrary: $\tilde{O}(n^4)$ queries
Posterior mean + Gaussian: $\tilde{O}(n^{2.5})$queries

Computationally bounded mec



- Suppose M is polynomial time
- Use public-key crypto to conceal T in tracing attack [Nissim Stemmer]
  - Public info: $pk_1, pk_2, \dots, pk_n$
  - $U = \{(i, sk_i) : i = 1, \dots, N\}$
  - $X = \{(i, sk_i) : i \in S\}$ 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

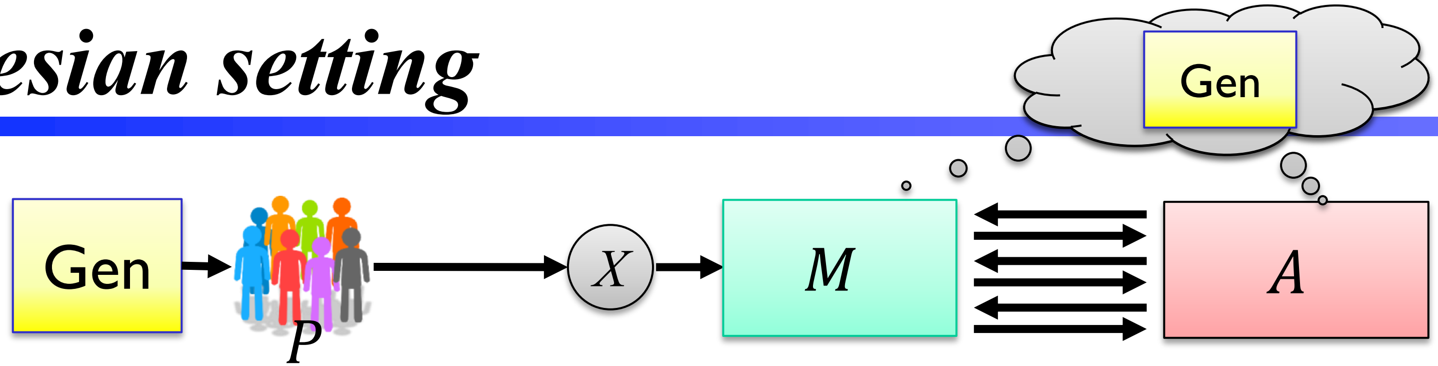
What can we say about variance?

- 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
    - Inferactive 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