Algorithmic Approaches to Preventing Overfitting in Adaptive Data Analysis

Part 2

Adam Smith

Boston University

Simons Institute workshop on adaptive data analysis
July 24, 2018
Part 2: Hiding the data

• Three related notions
  - Privacy
  - Algorithmic stability
  - Bounded information

• All three relate adaptive setting to execution on fresh data
  - Common idea: With limited information about the data, cannot overfit

• Larger goal: Prescriptive theory
  - Understand how to design algorithms to maximize data set’s long-term value
Adaptive Linear Queries

- Each query is a function $q: X \rightarrow [0,1]$
- Empirical answer
  $$q(X) = \frac{1}{n} \sum_{i} q(x_i)$$
- "Population answer"
  $$q(P) = \mathbb{E}_{Z \sim P}(q(Z))$$
- Answers have error $\alpha$ if $|a_i - q_i(P)| \leq \alpha \ (\forall i)$

**Examples**
- Contingency tables
- Classification error
- Optimization via gradient descent

[Dwork, Feldman, Hardt, Pitassi, Reingold, Roth 2015]
Adaptive Linear Queries

- Each query is a function \( q: \mathcal{X} \rightarrow [0,1] \)
- Empirical answer
  \[
  q(X) = \frac{1}{n} \sum_{i} q(x_i)
  \]
- “Population answer”
  \[
  q(P) = \mathbb{E}_{Z \sim P}(q(Z))
  \]
- Answers have error \( \alpha \) if
  \[
  |a_i - q_i(P)| \leq \alpha \quad (\forall i)
  \]
Adaptive Linear Queries

- Each query is a function \( q : \mathcal{X} \to [0,1] \)

- Empirical answer
  \[
  q(X) = \frac{1}{n} \sum_{i} q(x_i)
  \]

- “Population answer”
  \[
  q(P) = \mathbb{E}_{Z \sim P}(q(Z))
  \]

- Answers have error \( \alpha \) if
  \[
  |a_i - q_i(P)| \leq \alpha \quad (\forall i)
  \]

Nonadaptive queries

- Tracing queries [Steinke Ullman '15]
  \[
  \frac{\sqrt{\log k}}{\sqrt{n}}
  \]

- \( \frac{1}{\sqrt{n}} + \frac{\sqrt{k}}{n} \)

- \( \frac{4\sqrt{k}}{\sqrt{n}} \)

- \( \frac{k^{1/5}}{n^{2/5}} \)

Tracing queries [Steinke Ullman '15] $\frac{\sqrt{\log k}}{\sqrt{n}}$; Nonadaptive queries $\frac{1}{\sqrt{n}} + \frac{\sqrt{k}}{n}$; Adaptive queries $\frac{4\sqrt{k}}{\sqrt{n}}$; Population error bound $\frac{k^{1/5}}{n^{2/5}}$.
Adaptive Linear Queries

- Each query is a function $q : \mathcal{X} \to [0,1]$
- Empirical answer $q(X) = \frac{1}{n} \sum_i q(x_i)$
- “Population answer” $q(P) = \mathbb{E}_{Z \sim P}(q(Z))$
- Answers have error $\alpha$ if $|a_i - q_i(P)| \leq \alpha$ (\forall i)

Nonadaptive queries
$\log k / \alpha^2$ [Steinke Ullman ‘15]

Tracing queries
$1 / \alpha^2 + \sqrt{k} / \alpha$ [BNSSSU’16, RRST’16]

$\sqrt{k} / \alpha^2$ [DFHPRR’15]
Outline

• Privacy, Stability, Generalization: Pick Any Three
  ➢ “Stable algorithms cannot overfit”

• Applications to statistical queries
  ➢ “Transfer theorems” for stable algorithms

• Information and generalization
Differential Privacy

• Data set $x = (x_1, ..., x_n) \in D^n$
  - Domain $D$ can be numbers, categories, tax forms
  - Think of $x$ as fixed (not random)

• $A =$ randomized procedure
  - $A(x)$ is a random variable
  - Randomness might come from adding noise, resampling, etc.
Differential Privacy

- A thought experiment
  - Change one person’s data (or remove them)
  - Will the distribution on outputs change much?
Differential Privacy

$\mathbf{x}'$ is a neighbor of $\mathbf{x}$ if they differ in one data point.

**Definition:** $\mathbf{A}$ is $(\varepsilon, \delta)$-differentially private if, for all neighbors $\mathbf{x}, \mathbf{x}'$, for all subsets $\mathbf{S}$ of outputs

\[
\Pr(A(\mathbf{x}) \in \mathbf{S}) \leq e^{\varepsilon} \Pr(A(\mathbf{x}') \in \mathbf{S}) + \delta
\]
A is $\rho$-stable with respect to divergence $D$ if for all neighbors $x, x'$:

$$D(A(x), A(x')) \leq \rho$$

Here, $D$ could be KL, $\chi^2$, Renyi divergence, or other...

$x'$ is a neighbor of $x$ if they differ in one data point.

**Definition:** A is $(\epsilon, \delta)$-differentially private if, for all neighbors $x, x'$, for all subsets $S$ of outputs

$$\Pr(A(x) \in S) \leq e^\epsilon \Pr(A(x') \in S) + \delta$$
Why distributional stability?

With the right divergence, distributional stability…

- Is closed under processing by arbitrary analyst
  - Don’t need to understand how analyst works

- Degrades gracefully when algorithms are composed
  - If each $A_i$ is $(\epsilon_i, \delta_i)$-DP, then $B$ is $\approx (\epsilon \sqrt{k}, \delta k) – DP$
**Laplace Mechanism**

- Say we want to release a summary $f(x) \in \mathbb{R}^k$
  
  - e.g., proportion of diabetics: $x_i \in \{0,1\}$ and $f(x) = \frac{1}{n} \sum_i x_i$

- Simple approach: add noise to $f(x)$
  
  - How much noise is needed?
Laplace Mechanism

- Global Sensitivity: \( GS_f = \max_{\text{neighbors } x, x'} \| f(x) - f(x') \|_1 \)

- Example: \( GS_{\text{proportion}} = \frac{1}{n} \)
Laplace Mechanism

- Global Sensitivity: \( \text{GS}_f = \max_{\text{neighbors } x, x'} \| f(x) - f(x') \|_1 \)

- Example: \( \text{GS}_{\text{proportion}} = \frac{1}{n} \)

Theorem: \( A(x) = f(x) + \text{Lap} \left( \frac{\text{GS}_f}{\epsilon} \right) \) is \((\epsilon, 0)\)-differentially private.

- Laplace distribution \( \text{Lap}(\lambda) \) has density
  \[ h(y) \propto e^{-|y|/\lambda} \]
- Changing one point translates curve
A rich algorithmic field

Noise addition

Exponential sampling

\[ Y \sim p(y|x) \propto \exp(\epsilon \cdot \text{quality}(y, x)) \]

Local perturbation

\[ x_1, x_2, \ldots, x_n \rightarrow Q_1, Q_2, \ldots, Q_n \rightarrow \text{Untrusted aggregator} \rightarrow A \]
Outline

• Privacy, Stability, Generalization: Pick Any Three
  ➢ “Stable algorithms cannot overfit”

• Applications to statistical queries
  ➢ “Transfer theorems” for stable algorithms

• Information and generalization
Why distributional stability?

- Implies that the analyst “cannot overfit”. Suppose:
  - Analyst chooses $P$
  - Algorithm produces output $a = A(X)$
  - Analyst selects a statistical query $q_\alpha : \rightarrow [0,1]$

\[
\text{Score} = |q_\alpha(X) - q_\alpha(P)| \\
\approx |q_\alpha(X) - q_\alpha(X')|
\]

**Meta-Theorem** [DFHPRR, …]:
If $A$ is $\rho$-stable w.r.t. $D$, then: $\forall P$, $\forall$ analysts:

\[
\text{Score} \leq f(\rho, D)
\]

with high probability.
Generalization Lemmas

\[ \text{Score} = |q_a(X) - q_a(P)| \quad \text{where } a = A(X) \]
\[ \approx |q_a(X) - q_a(X')| \]

- \((\varepsilon, \delta)\)-DP \(\Rightarrow\) score = \(O(\varepsilon)\) with probability \(\approx 1 - e^{-\varepsilon^2 n - \delta/\varepsilon}\) \[\text{[DFHPRR '15, BNSSSU '16]}\]

- \(\varepsilon\)-TV stable \(\Rightarrow\) \(E(\text{score}) = \varepsilon\) \[\text{[McSherry ??]}\]

- \(\varepsilon^2\)-KL stable \(\Rightarrow\) \(\sqrt{E(\text{score}^2)} = O(\varepsilon)\) \[\text{[Russo-Zou '15, WangLeiFienberg'16]}\]

- \(\varepsilon^2\)-"zCDP" \(\Rightarrow\) score = \(O(\varepsilon)\) with high prob. \[\text{[Bun, Dwork, Rothblum, Steinke]}\]
Proof idea: Stability

• **Lemma**: If $A$ is $\epsilon$-TV stable, then for all distributions $P$:
  
  $$
  E_{X \sim P^n} \left( q_a(X) - q_a(P) \right) \leq \epsilon
  $$

• **Proof**:
  
  ➢ Fix distribution $P$
  ➢ Compare distributions on two triples
    • $(\vec{X}, i, A(\vec{X}))$ and $(\vec{X}, i, A(\vec{X}_{-i}, \vec{x}))$ where $x_1, \ldots, x_n, \vec{x} \sim P$ are i.i.d.
  ➢ Observation: These have total variation distance $\leq \epsilon$.
    • Expectations of bounded functions are about the same
  ➢ Consider the bounded function $f(\vec{x}, i, y) = q_y(x_i)$ where $q_y$ is the query selected by analyst on output
  ➢ Now we have
    
    $$
    E \left( f(\vec{X}, i, A(\vec{X})) \right) = E(q_a(\vec{x}))
    $$
    
    $$
    E \left( f(\vec{X}, i, A(\vec{X}_{-i}, \vec{x})) \right) = E(q_a(P))
    $$
  ➢ So $E(q_a(X) - q_a(P)) \leq \epsilon$

• Need a bit more work to get $E(score) \leq \epsilon$
High-Probability Bounds

• To get subgaussian concentration, need stronger guarantees than TV or KL stability
  ➢ $(\epsilon, \delta)$-differential privacy currently the best

• Idea [Nissim-Stemmer]:
  ➢ Run $t \approx 1/\delta$ copies of the game with independent data sets
    • If analyst succeeds with probability $\delta$, then with constant probability
      one of the copies produced a query that overfit
  ➢ Use a differentially private algorithm to choose copy with “worst” error
  ➢ Argue that composed algorithm…
    • Is differentially private [easy]
    • Should not be able to overfit to any of the $t$ data sets [subtle]
Outline

• Privacy, Stability, Generalization: Pick Any Three
  ➢ “Stable algorithms cannot overfit”

• Applications to statistical queries
  ➢ “Transfer theorems” for stable algorithms

• Information and generalization
Adaptive Linear Queries

- Each query is a function $q : X \rightarrow [0,1]$
- Empirical answer
  \[ q(X) = \frac{1}{n} \sum_{i} q(x_i) \]
- "Population answer"
  \[ q(P) = \mathbb{E}_{Z \sim P} (q(Z)) \]
- Answers have error $\alpha$ if
  \[ |a_i - q_i(P)| \leq \alpha \quad (\forall i) \]

**Nonadaptive queries**
\[
\frac{\sqrt{\log k}}{\sqrt{n}}
\]
- Tracing queries
  - Steinke Ullman '15
  \[
  \frac{1}{\sqrt{n}} + \frac{\sqrt{k}}{n}
  \]
- [BNSSSU'16, RRST'16]
- [DFHPRR'15]

- [BNSSSU'16, RRST'16]
  \[
  \frac{4\sqrt{k}}{\sqrt{n}}
  \]
  \[
  \frac{k^{1/5}}{n^{2/5}}
  \]
“Transfer” Theorem

• The generalization lemmas connect accuracy on the population with sample accuracy.

• We say $A$ is $(\alpha, \beta)$ sample-accurate if, for all data sets $x$,
  \[ \max_i |a_i - q_i(x)| \leq \alpha \]
  with probability \( \geq 1 - \beta \).

• **Theorem [BNSSSSU]**: If $A$ is $(\epsilon, \delta)$-DP and $(\alpha, \beta)$-sample accurate, then
  \[ \max_i |a_i - q_i(P)| \leq O(\alpha + \epsilon) \]
  with probability \( \geq 1 - \beta - \delta/\epsilon \).

• Similar theorems possible for weaker stability notions

• Proof relies on “right” way to handle many rounds
From 2 to k stages: Induction [DFHPRR’15]

- Apply overfitting lemma at each round
  - Probability of overfitting adds up over rounds
**“Monitor Argument”** [BNSSSSU’16]

- **Stronger bounds**
- **Generalizes beyond linear queries**

1. Find $i^*$
   $$i^* = \arg\max_i |a_i - q_i(P)|$$
2. Return $q_i^*$

**Observation:**
$$\epsilon \geq Score(A) \geq \max_i |a_i - q_i(P)| - \alpha$$
Application 1: Worst-case queries

• One can answer an arbitrary sequence of $k$ adaptively chosen statistical queries such that (w.h.p.)

$$\max_i |a_i - q_i(P)| = \tilde{O}\left(\frac{4\sqrt{k}}{\sqrt{n}}\right)$$

➢ Alternatively, for error $\alpha$, a sufficient sample size is

$$n = \tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$$

• Algorithm: On each query, add Laplace (or Gaussian) noise with standard deviation

$$\frac{4\sqrt{k}}{\sqrt{n}}$$
Adding noise to many queries

• Suppose we have \( k \) statistical queries \( q_1, \ldots, q_k \)

• **Lemma:** There is an \((\varepsilon, \delta)\)-differentially private algorithm that answers each query with sample error

\[
\max_i |a_i - q_i(x)| = O_P \left( \frac{\sqrt{k}}{\varepsilon n} \cdot \sqrt{\ln(k) \ln(1/\delta)} \right)
\]

• Run Laplace mechanism \( k \) times,
  - with parameter \( \varepsilon' \approx \varepsilon/\sqrt{k} \)
  - then apply composition theorems

• **Corollary** (via Transfer Theorem): If \( X \sim P^n \), then

\[
\max_i |a_i - q_i(P)| = \tilde{O} \left( \frac{\sqrt{k}}{\varepsilon n} + \varepsilon \right) = \tilde{O} \left( \frac{4\sqrt{k}}{\sqrt{n}} \right).
\]
Application 2: Reusable Holdout [DFHPRR]

- Recall from part 1: we can answer $k$ queries with error nearly independent of $k$
  - Use “dirty” set $S$ to generate guesses, and “clean” set $C$ to verify.
  - Algorithm: answer only those queries where $|q_i(X_S) - q_i(X_C)| > T$ for some $T$
  - Error is $T + \tilde{O}\left(\frac{\sqrt{w \log k}}{\sqrt{n}}\right)$

- New version: add noise each time you compare to threshold
  - Obtain error $T + \tilde{O}\left(\frac{(w \log k)^{1/4}}{\sqrt{n}}\right)$
Sparse vector mechanism

• Suppose we have \( k \) statistical queries \( q_1, \ldots, q_k \)
  – Each asks for the average of a \([0,1]\) function over the data
  – Posed adaptively

• We want to know which queries exceed a threshold \( T \)
  – E.g. which queries are way above a guessed value
  – Can we pay only for the number of queries above threshold?

• Sparse Vector Mechanism* \( (x, q_1, q_2, \ldots) \)
  – Flags = 0
  – While(Flags < w):
    • Receive next query \( q_i \)
    • If \( \left( q_i(x) + \text{Lap}\left( \frac{1}{n\epsilon'} \right) > T \right) \):
      – Answer “above threshold”
      – Flags \( \leftarrow \) Flags + 1
    • Else
      – Answer “below threshold”

Theorem*: For \( \epsilon' \approx \frac{\epsilon}{\sqrt{w \ln(1/\delta)}} \),
Sparse Vector is
• \((\epsilon, \delta)\) -DP
• Correct w.h.p. for all \( i \) s.t.
  \[ |q_i(x) - T| \geq \Omega \left( \frac{\sqrt{w \ln(1/\delta) \ln k}}{n\epsilon} \right) \]

* Actual algorithm also randomizes \( T \)
Similar applications

• Median mechanism
  ➢ Compression analysis $\tilde{O} \left( \frac{\log |x| \cdot \log k}{n} \right)^{1/4}$
  ➢ Stability-based: $\tilde{O} \left( \frac{(\log k)^{1/2} (\log |x|)^{1/6}}{\sqrt{n}} \right)$

• Ladder algorithm [Hardt17]
  ➢ Compression analysis $n = \frac{\log k}{\alpha^3}$
  ➢ Stability-based: $n = \frac{(\log k)^{1.5}}{\alpha^{2.5}}$
Outline

• Privacy, Stability, Generalization: Pick Any Three
  ➢ “Stable algorithms cannot overfit”

• Applications to statistical queries
  ➢ “Transfer theorems” for stable algorithms

• Information and generalization
Information and Overfitting

- Look at **information** in $Y = A(X)$ about $X$
- Several measures based on **odds ratio**

$$I_{x,y} = \log \left( \frac{\Pr(A(X) = y \mid X = x)}{\Pr(A(X) = y)} \right)$$

- Mutual information: expectation of $I_{x,y}$
- Max information: high-probability bound on $I_{x,y}$
- Min-entropy leakage: $\mathbb{E}_{y \sim Y} \left( \sup_x I_{x,y} \right)$

[DFHPRR, Russo-Zou, RRST, Xu-Raginsky,...]

Strongest guarantees
Information and Overfitting

- Look at **information** in $Y = A(X)$ about $X$
- Several measures based on **odds ratio**

$$I_{x,y} = \frac{\Pr(A(X) = y \mid X = x)}{\Pr(A(X) = y)}.$$  

**Meta-Lemma:** score $\lesssim \sqrt{\text{information} / n}$

**Theorem:** If $A$ is $(\epsilon, \delta)$-DP*, then $\max - \text{info} \lesssim \epsilon^2 n$.

**Theorem:** If $A$ is $\ell$-compressible, then $\max - \text{info} \lesssim \ell$.  

[DFHPRR, Russo-Zou, RRST, Xu-Raginsky, …]
From information to hypothesis testing

• Consider adaptive hypothesis selection: analyst makes a conjecture $H_0$ about $P$, and chooses a test $T$ such that $\Pr(T(X) = 1|P \in H_0, X \sim P^n) \leq p_0$

• The max information is

$$I_\infty(X; A(X)) = \max_{x,y} \log \frac{\Pr(A(x) = y|X = x)}{\Pr(A(x) = y)}$$

• Observation: If $I_\infty(X; A(X)) \leq k$, then

$$\Pr(T(X) = 1|P \in H_0, X \sim P^n, T = A(X)) \leq 2^k p_0.$$  

• Other measures of information yield more complex relationships

  ➢ Not yet well explored [Russo-Zou’15, RogersRST16, S’17]
• Privacy, Stability, Generalization: Pick Any Three
  ➢ “Stable algorithms cannot overfit”

• Applications to statistical queries
  ➢ “Transfer theorems” for stable algorithms

• Information and generalization
Conclusions

- Adaptive analysis is everywhere
  - “All inference” is selective
- We can get nontrivial results for arbitrary analyst behavior
  - Accuracy/power guarantees
  - Results are (essentially) tight
  - Information and stability play key roles
- Current theory most useful for
  - Many queries
  - Statistical queries
- Not covered
  - Lower bounds on accuracy (and open problems)
  - Concrete bounds (see talks by Feldman and Thakkar)
  - Accuracy as a good: allocating costs (fairly?)
  - Models of “benign” analyst (see my second talk)
  - Adaptive hypothesis testing
- Lecture notes for Penn-BU course at http://adaptive-dataanalysis.com