Sampling Correctors

If the data don't fit the theory, change the data.

Clément Canonne
Columbia University

January 14, 2016

Joint work with Themis Gouleakis (MIT) and Ronitt Rubinfeld (MIT).

What is data?
(Not a random question)

163	205	199	77	199	245	43	149	92	40
166	116	151	97	72	114	240	77	249	54
194	47	188	100	249	9	145	136	111	61
161	124	242	42	158	110	236	33	127	204
201	124	52	234	166	43	73	98	121	54
223	32	203	12	18	252	224	32	20	49
127	35	195	203	68	214	166	163	124	52
241	215	251	78	42	192	86	10	230	242
181	77	66	160	59	221	182	159	104	252
40	220	251	35	236	49	120	206	181	164
213	106	99	148	142	90	104	55	164	133

What is data?

(Not a random question)

140	237	98	230	243	40	77	101	33	192
154	40	46	198	173	133	222	196	42	193
109	129	43	11	115	132	213	29	71	131
144	72	156	125	95	132	105	63	189	197
185	138	237	104	124	113	38	63	196	141
205	109	17	188	201	82	132	72	176	254
208	62	236	226	223	220	219	189	40	145
239	126	69	178	56	36	63	79	84	178
201	2	225	62	179	243	191	118	134	90
209	178	146	252	196	221	130	137	124	124
27	58	114	139	215	207	172	3	178	205

What is data?

(Not a random question)

194	62	174	34	52	93	241	78	163	48
67	111	115	169	99	223	174	72	68	150
31	254	38	225	193	90	33	174	27	53
230	160	56	18	117	11	205	81	87	193
93	255	104	70	166	242	232	219	104	81
6	119	17	80	1	87	106	25	93	185
234	93	173	210	172	102	95	50	198	28
166	128	111	249	124	90	125	110	25	31
148	234	47	73	71	67	4	130	165	39
227	227	5	23	70	226	101	98	147	54
255	39	246	149	70	211	144	204	176	95

What is data?

(Not a random question)

101	61	105	246	143	142	36	211	10	104
119	8	149	212	223	251	114	66	38	182
54	4	91	183	152	133	16	37	35	195
190	116	171	65	9	62	160	99	155	108
23	130	223	64	239	159	70	247	46	254
63	83	191	109	67	170	2	107	5	242
137	162	127	54	241	182	86	25	137	197
203	215	188	91	155	227	251	23	53	226
162	214	14	137	95	165	188	243	179	249
203	191	171	32	119	240	126	40	193	5
213	184	196	147	93	3	184	90	173	88

What is data?
(Not a random question)

Datapoints $\equiv$ Independent *samples* from a distribution:

$$x_1, x_2, \ldots, x_m \sim D$$

which can only be accessed via this "sampling oracle."

What is data?

(Not a random question)

Datapoints $\equiv$ Independent *samples* from a distribution:

$$x_1, x_2, \ldots, x_m \sim D$$

which can only be accessed via this "sampling oracle."

What can be known about D?

Distributions: testing, learning and. . . ?

Challenges and paradigms

Distribution D over domain of size n – but n is ginormous.

Distributions: testing, learning and. . . ?

Challenges and paradigms

Distribution D over domain of size n – but n is ginormous.

Learning distributions:

D is promised to belong to some class $\mathcal{P}$: using $o(n)$ samples, output a hypothesis $\hat{D}$ that approximates D.

DISTRIBUTIONS: TESTING, LEARNING AND. . . ?

CHALLENGES AND PARADIGMS

Distribution D over domain of size n – but n is ginormous.

Learning distributions:

D is promised to belong to some class $\mathcal{P}$: using $o(n)$ samples, output a hypothesis $\hat{D}$ that approximates D.

Testing distributions:

using $o(n)$ samples – ideally $O(1)$, decide whether D belongs to some class $\mathcal{P}$, or is far from it.

Distributions: testing, learning and...?

Challenges and paradigms

Distribution D over domain of size n – but n is ginormous.

Learning distributions:

D is promised to belong to some class $\mathcal{P}$: using $o(n)$ samples, output a hypothesis $\hat{D}$ that approximates D.

Testing distributions:

using $o(n)$ samples – ideally $O(1)$, decide whether D belongs to some class $\mathcal{P}$, or is far from it.

does that cover everything?

Motivation

A slide with text

Often, the distribution on the data has particular, useful structure that algorithms can exploit (monotone pmf, uniform distribution, independent components…)

MOTIVATION

A SLIDE WITH TEXT

Often, the distribution on the data has particular, useful structure that algorithms can exploit (monotone pmf, uniform distribution, independent components...)

But in many situations, sample data comes from noisy or imperfect sources, tampering with these properties.

Motivation

A slide with text

Often, the distribution on the data has particular, useful structure that algorithms can exploit (monotone pmf, uniform distribution, independent components…)

But in many situations, sample data comes from noisy or imperfect sources, tampering with these properties.

Can we still exploit the structure the distribution should have had?

MOTIVATION

A SLIDE WITH PICTURES

Figure : Whooping! *"Some data sets, however, may contain both systematic and random errors in the recorded location of the species."* [Hefley et al., 2014]

Motivation

A slide with pictures

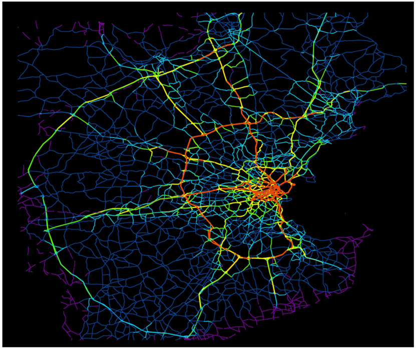


Figure : Analyzing the traffic when some sensors went haywire?

Motivation

A slide with pictures

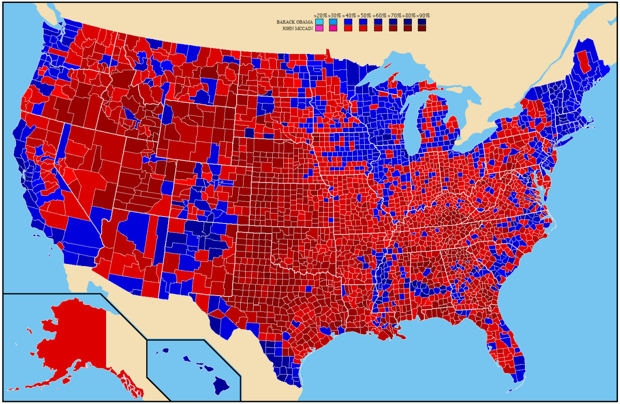


Figure : “We might be missing some of the votes from state *blah*.”

Motivation

A slide with pictures

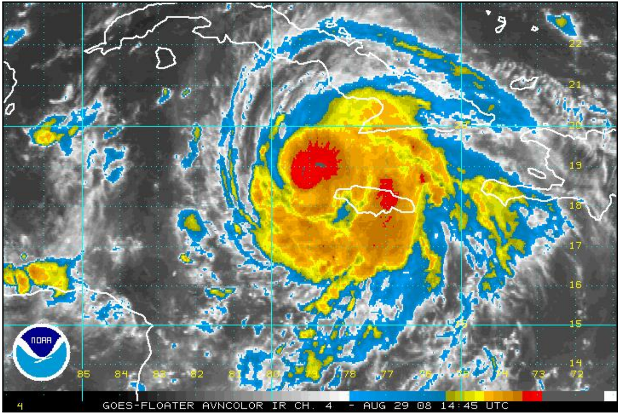


Figure : Sensors can go off – and do.

FROM THERE...

How to address these problems?

FROM THERE. . .

~~How to address these problems?~~

How to model these problems?

NEITHER LEARNING NOR TESTING

"AND NOW, FOR SOMETHING COMPLETELY DIFFERENT."

Correcting distributions

A general methodology

Fix a *specific* property $\mathcal{P}$ of distributions.
(application-dependent)

- independent samples from a D promised to be ε-close to $\mathcal{P}$
- want independent samples from $\tilde{D}$ which:
  - has the property: $\tilde{D} \in \mathcal{P}$;
  - remains faithful to the data: $\mathrm{d}_{\mathrm{TV}}\big(\tilde{D}, D\big) = O(\varepsilon)$.

Correcting distributions

A general methodology

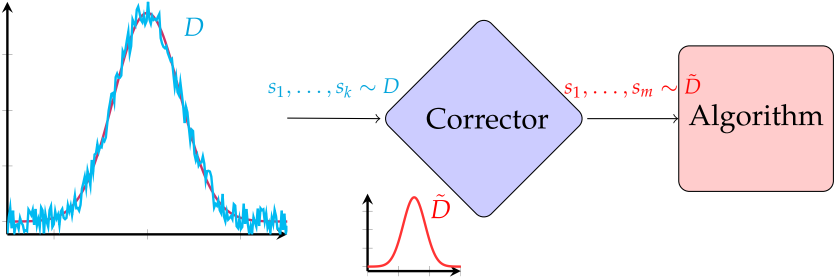
Fix a *specific* property $\mathcal{P}$ of distributions.
(application-dependent)

- independent samples from a D promised to be ε-close to $\mathcal{P}$
- want independent samples from $\tilde{D}$ which:
  - has the property: $\tilde{D} \in \mathcal{P}$;
  - remains faithful to the data: $\mathrm{d}_{\mathrm{TV}}\left(\tilde{D}, D\right) = O(\varepsilon)$.

Similar in spirit to the "local filters" for functions [Ailon et al., 2008, Saks and Seshadhri, 2010, Jha and Raskhodnikova, 2011, Bhattacharyya et al., 2012].

Correcting distributions

A general methodology



Correcting distributions

Challenges

Correcting distributions

Challenges

sample rate How many samples of D per sample of $\tilde{D}$?

Correcting distributions

Challenges

sample rate How many samples of D per sample of $\tilde{D}$?

randomness How much *extra* randomness is needed?

Rest of the talk

A glimpse at results.

1. Connections to learning and testing
2. Randomness scarcity: no coins of our own (uniformity correction)
3. Beating the learning approach: the case of monotonicity

CONNECTIONS

What does the existence of sampling correctors imply for learnability or testability?

CONNECTIONS

- Agnostic learner $\rightarrow$ Sample corrector

CONNECTIONS

- Agnostic learner $\rightarrow$ Sample corrector
- Sample corrector + distance approximator + tester $\rightarrow$ *tolerant* tester

CONNECTIONS

- Agnostic learner → Sample corrector
- Sample corrector + distance approximator + tester → *tolerant* tester
- Sample corrector + learner → *agnostic* learner

CONNECTIONS

- Agnostic learner $\to$ Sample corrector
- Sample corrector + distance approximator + tester $\to$ *tolerant* tester
- Sample corrector + learner $\to$ *agnostic* learner

Instantiate: get weakly tolerant monotonicity testers for k-modal.

Randomness scarcity

The case of uniformity

Coins don't come cheap

Can we leverage the inherent randomness of the data to use only few random coins of our own?

Sample complexity

The case of monotonicity

Beating the learning approach

Can we correct a distribution efficiently, without having to learn it?

Sample complexity

The case of monotonicity: some results

- Can correct *really* small error with rate $O(1)$

SAMPLE COMPLEXITY

THE CASE OF MONOTONICITY: SOME RESULTS

- Can correct *really* small error with rate $O(1)$
- Can correct with rate $O(\sqrt{\log n})$ *with stronger (CDF) queries*

Sample complexity

The case of monotonicity: some results

- Can correct *really* small error with rate $O(1)$
- Can correct with rate $O(\sqrt{\log n})$ *with stronger (CDF) queries*
- Can correct *specific* types of errors with rate $O(1)$

Sample complexity

The case of monotonicity: some results

- Can correct *really* small error with rate $O(1)$
- Can correct with rate $O(\sqrt{\log n})$ *with stronger (CDF) queries*
- Can correct *specific* types of errors with rate $O(1)$

…but constant error with rate $o(\log n)$ seems ruled out

Conclusion

- $o(\log n/\varepsilon^3)$ corrector for monotonicity?
- what about independence?
- leveraging the connections to (re)derive learning and testing upper and lower bounds?

Conclusion

- $o(\log n/\varepsilon^3)$ corrector for monotonicity?
- what about independence?
- leveraging the connections to (re)derive learning and testing upper and lower bounds?

Meta question

Which properties $\mathcal{P}$ can we correct efficiently – and which ones arise in which scenarios?

Thank you.

N. Ailon, B. Chazelle, S. Comandur, and D. Liu. Property-preserving data reconstruction. *Algorithmica*, 51(2):160–182, 2008.

A. Bhattacharyya, E. Grigorescu, M. Jha, K. Jung, S. Raskhodnikova, and D. P. Woodruff. Lower bounds for local monotonicity reconstruction from transitive-closure spanners. *SIAM Journal on Discrete Mathematics*, 26(2):618–646, 2012.

T. J. Hefley, D. M. Baasch, A. J. Tyre, and E. E. Blankenship. Correction of location errors for presence-only species distribution models, 2014. ISSN 2041-210X.

M. Jha and S. Raskhodnikova. Testing and reconstruction of Lipschitz functions with applications to data privacy. 2011.

M. Saks and C. Seshadhri. Local monotonicity reconstruction. *SIAM Journal on Computing*, 39(7):2897–2926, 2010.