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)

<table>
<thead>
<tr>
<th>Background</th>
<th>Model</th>
<th>(A glimpse of) results</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>163</td>
<td>205</td>
<td>199</td>
<td>77</td>
</tr>
<tr>
<td>199</td>
<td>245</td>
<td>43</td>
<td>149</td>
</tr>
<tr>
<td>199</td>
<td>245</td>
<td>43</td>
<td>149</td>
</tr>
<tr>
<td>163</td>
<td>205</td>
<td>199</td>
<td>77</td>
</tr>
<tr>
<td>166</td>
<td>116</td>
<td>151</td>
<td>97</td>
</tr>
<tr>
<td>97</td>
<td>72</td>
<td>114</td>
<td>240</td>
</tr>
<tr>
<td>240</td>
<td>77</td>
<td>249</td>
<td>54</td>
</tr>
<tr>
<td>194</td>
<td>47</td>
<td>188</td>
<td>100</td>
</tr>
<tr>
<td>249</td>
<td>9</td>
<td>145</td>
<td>136</td>
</tr>
<tr>
<td>136</td>
<td>111</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>124</td>
<td>42</td>
<td>158</td>
</tr>
<tr>
<td>242</td>
<td>42</td>
<td>158</td>
<td>110</td>
</tr>
<tr>
<td>236</td>
<td>33</td>
<td>127</td>
<td>204</td>
</tr>
<tr>
<td>201</td>
<td>124</td>
<td>52</td>
<td>234</td>
</tr>
<tr>
<td>234</td>
<td>166</td>
<td>43</td>
<td>73</td>
</tr>
<tr>
<td>73</td>
<td>98</td>
<td>121</td>
<td>54</td>
</tr>
<tr>
<td>223</td>
<td>32</td>
<td>203</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>252</td>
<td>224</td>
</tr>
<tr>
<td>224</td>
<td>32</td>
<td>20</td>
<td>49</td>
</tr>
<tr>
<td>127</td>
<td>35</td>
<td>195</td>
<td>203</td>
</tr>
<tr>
<td>203</td>
<td>68</td>
<td>214</td>
<td>166</td>
</tr>
<tr>
<td>214</td>
<td>166</td>
<td>163</td>
<td>124</td>
</tr>
<tr>
<td>124</td>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>241</td>
<td>215</td>
<td>251</td>
<td>78</td>
</tr>
<tr>
<td>251</td>
<td>78</td>
<td>42</td>
<td>192</td>
</tr>
<tr>
<td>192</td>
<td>86</td>
<td>10</td>
<td>230</td>
</tr>
<tr>
<td>230</td>
<td>242</td>
<td></td>
<td></td>
</tr>
<tr>
<td>181</td>
<td>77</td>
<td>66</td>
<td>160</td>
</tr>
<tr>
<td>66</td>
<td>59</td>
<td>221</td>
<td>182</td>
</tr>
<tr>
<td>221</td>
<td>182</td>
<td>159</td>
<td>104</td>
</tr>
<tr>
<td>104</td>
<td>252</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>220</td>
<td>251</td>
<td>35</td>
</tr>
<tr>
<td>251</td>
<td>35</td>
<td>236</td>
<td>49</td>
</tr>
<tr>
<td>236</td>
<td>49</td>
<td>120</td>
<td>206</td>
</tr>
<tr>
<td>120</td>
<td>206</td>
<td>181</td>
<td>164</td>
</tr>
<tr>
<td>181</td>
<td>164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>213</td>
<td>106</td>
<td>99</td>
<td>148</td>
</tr>
<tr>
<td>99</td>
<td>148</td>
<td>142</td>
<td>90</td>
</tr>
<tr>
<td>142</td>
<td>90</td>
<td>104</td>
<td>55</td>
</tr>
<tr>
<td>104</td>
<td>55</td>
<td>164</td>
<td>133</td>
</tr>
</tbody>
</table>
WHAT IS DATA?
(Not a random question)
What is data?

(Not a random question)

<table>
<thead>
<tr>
<th>Background</th>
<th>Model</th>
<th>(A glimpse of) results</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>194 62 174 34 52 93 241 78 163 48</td>
<td>67 111 115 169 99 223 174 72 68 150</td>
<td>31 254 38 225 193 90 33 174 27 53</td>
<td>230 160 56 18 117 11 205 81 87 193</td>
</tr>
<tr>
<td>93 255 104 70 166 242 232 219 104 81</td>
<td>6 119 17 80 1 87 106 25 93 185</td>
<td>234 93 173 210 172 102 95 50 198 28</td>
<td>166 128 111 249 124 90 125 110 25 31</td>
</tr>
<tr>
<td>148 234 47 73 71 67 4 130 165 39</td>
<td>227 227 5 23 70 226 101 98 147 54</td>
<td>255 39 246 149 70 211 144 204 176 95</td>
<td></td>
</tr>
</tbody>
</table>
**WHAT IS DATA?**
*(Not a random question)*

<table>
<thead>
<tr>
<th>Background</th>
<th>Model</th>
<th>(A glimpse of) results</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>101</strong></td>
<td><strong>61</strong></td>
<td><strong>105</strong></td>
<td><strong>246</strong></td>
</tr>
<tr>
<td><strong>119</strong></td>
<td><strong>8</strong></td>
<td><strong>149</strong></td>
<td><strong>212</strong></td>
</tr>
<tr>
<td><strong>54</strong></td>
<td><strong>4</strong></td>
<td><strong>91</strong></td>
<td><strong>183</strong></td>
</tr>
<tr>
<td><strong>190</strong></td>
<td><strong>116</strong></td>
<td><strong>171</strong></td>
<td><strong>65</strong></td>
</tr>
<tr>
<td><strong>23</strong></td>
<td><strong>130</strong></td>
<td><strong>223</strong></td>
<td><strong>64</strong></td>
</tr>
<tr>
<td><strong>63</strong></td>
<td><strong>83</strong></td>
<td><strong>191</strong></td>
<td><strong>109</strong></td>
</tr>
<tr>
<td><strong>137</strong></td>
<td><strong>162</strong></td>
<td><strong>127</strong></td>
<td><strong>54</strong></td>
</tr>
<tr>
<td><strong>203</strong></td>
<td><strong>215</strong></td>
<td><strong>188</strong></td>
<td><strong>91</strong></td>
</tr>
<tr>
<td><strong>162</strong></td>
<td><strong>214</strong></td>
<td><strong>14</strong></td>
<td><strong>137</strong></td>
</tr>
<tr>
<td><strong>203</strong></td>
<td><strong>191</strong></td>
<td><strong>171</strong></td>
<td><strong>32</strong></td>
</tr>
<tr>
<td><strong>213</strong></td>
<td><strong>184</strong></td>
<td><strong>196</strong></td>
<td><strong>147</strong></td>
</tr>
</tbody>
</table>
**WHAT IS DATA?**

*(Not a random question)*

Datapoints ≡ 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 \text{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...)*
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*

![Map of the United States with color-coded states](image)

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 \( \varepsilon \)-close to \( \mathcal{P} \)
- want independent samples from \( \tilde{D} \) which:
  - has the property: \( \tilde{D} \in \mathcal{P} \);
  - remains faithful to the data: \( d_{TV}(\tilde{D}, D) = O(\varepsilon) \).
CORRECTING DISTRIBUTIONS
A general methodology

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

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

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

\[ s_1, \ldots, s_k \sim D \]
\[ s_1, \ldots, s_m \sim \tilde{D} \]
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
What does the existence of sampling correctors imply for learnability or testability?
CONNECTIONS

- Agnostic learner → Sample corrector
CONNECTIONS

- Agnostic learner → Sample corrector
- Sample corrector + distance approximator + tester → tolerant tester
CONNECTIONS

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

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

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

THE CASE OF UNIFORMITY

COSINS 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.


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.