Sampling Correctors

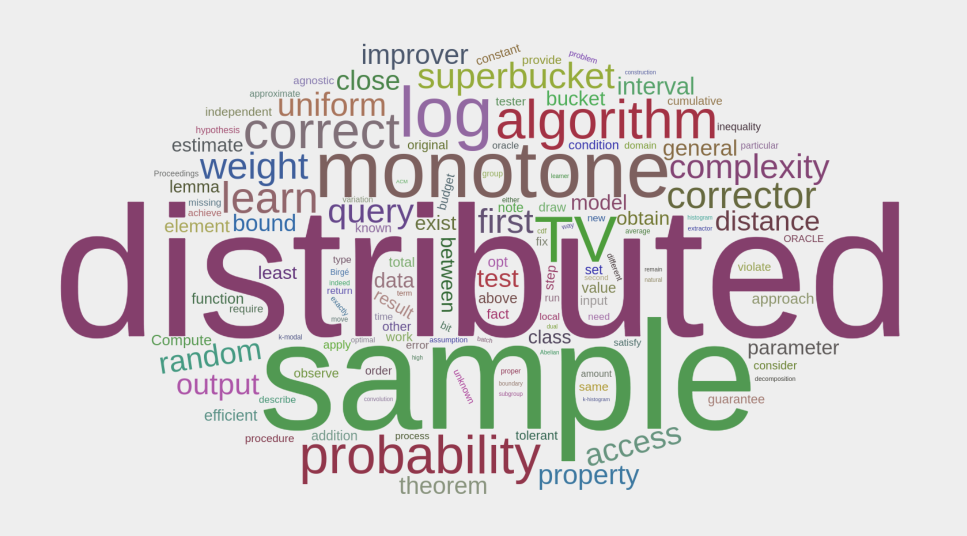
Clément L. Canonne, Themis Gouleakis, Ronitt Rubinfeld

The 7th Annual Innovations in Theoretical Computer Science conference (ITCS'16)

Speaker: Joseph Chuang-Chieh Lin

Institute of Information Science
Academia Sinica
Taiwan

28 April 2017


distributed
sample
monotone
algorithm
log
probability
correct
superbucket
improver
constant
provide
problem
construction
agnostic
close
interval
approximate
uniform
bucket
tester
cumulative
independent
inequality
hypothesis
estimate
original
oracle
condition
domain
general
particular
weight
complexity
Proceedings
ACM
group
learner
lemma
learn
budget
corrector
missing
variation
either
note
draw
model
achieve
query
histogram
element
bound
exist
first
TV
new
obtain
distance
known
cdf
way
extractor
average
fix
ORACLE
type
total
opt
different
violate
least
Birgé
data
test
step
set
remain
indeed
second
natural
return
between
value
term
result
above
run
input
approach
function
require
exactly
move
time
fact
local
need
other
bit
dual
class
k-modal
optimal
work
assumption
batch
Compute
apply
error
satisfy
parameter
random
high
Abelian
consider
observe
order
unknown
proper
amount
decomposition
output
boundary
same
describe
convolution
subgroup
k-histogram
guarantee
efficient
addition
process
tolerant
procedure
access
theorem
property

Outline

Motivations

- Data consisting of samples from distributions has reliability issues.
- If you know that the uncorrupted distribution is Gaussian, it would be natural to *correct* the samples to the nearest Gaussian.
- How do you correct the samples if you do NOT know much about the original uncorrupted distribution?

Motivations

- Data consisting of samples from distributions has reliability issues.
- If you know that the uncorrupted distribution is Gaussian, it would be natural to *correct* the samples to the nearest Gaussian.
- How do you correct the samples if you do NOT know much about the original uncorrupted distribution?

Motivations

- Data consisting of samples from distributions has reliability issues.
- If you know that the uncorrupted distribution is Gaussian, it would be natural to *correct* the samples to the nearest Gaussian.
- How do you correct the samples if you do NOT know much about the original uncorrupted distribution?

Contribution in general

- A methodology based on using *known structural properties* of the distribution to design sampling correctors which "correct" the sample data.
- **Question:** How best one can output samples of a distribution such that
  - on one hand, the structural properties are restored,
  - on the other hand, the corrected distribution, say $\tilde{D}$ is close to the original distribution, say D.
- We wish to optimize the two parameters:
  - # samples of D needed to output samples of $\tilde{D}$;
  - # additional truly random bits needed to output samples of $\tilde{D}$.

  ⋆ For any property $\mathcal{P}$, can one achieve improved query complexity in terms of these parameters over the use of the naïve learning approach for $\mathcal{P}$?

Contribution in general

- A methodology based on using *known structural properties* of the distribution to design sampling correctors which "correct" the sample data.
- **Question:** How best one can output samples of a distribution such that
  - on one hand, the structural properties are restored,
  - on the other hand, the corrected distribution, say $\tilde{D}$ is close to the original distribution, say D.
- We wish to optimize the two parameters:
  - $\#$ samples of D needed to output samples of $\tilde{D}$;
  - $\#$ additional truly random bits needed to output samples of $\tilde{D}$.

  ⋆ For any property $\mathcal{P}$, can one achieve improved query complexity in terms of these parameters over the use of the naïve learning approach for $\mathcal{P}$?

Contribution in general

- A methodology based on using *known structural properties* of the distribution to design sampling correctors which "correct" the sample data.
- **Question:** How best one can output samples of a distribution such that
  - on one hand, the structural properties are restored,
  - on the other hand, the corrected distribution, say $\tilde{D}$ is close to the original distribution, say D.
- We wish to optimize the two parameters:
  - $\#$ samples of D needed to output samples of $\tilde{D}$;
  - $\#$ additional truly random bits needed to output samples of $\tilde{D}$.

  ⋆ For any property $\mathcal{P}$, can one achieve improved query complexity in terms of these parameters over the use of the naïve learning approach for $\mathcal{P}$?

Sampling Corrector

- $\mathcal{P}$: a fixed and given distributions on Ω.
- A distribution D over $[n]$, $d_{TV}(D, \mathcal{P}) \leq \epsilon$.

An (ϵ, ϵ_1)-sampling corrector for $\mathcal{P}$ is a randomized algorithm which is given

- $\epsilon, \epsilon_1 \in (0, 1]$ s.t. $\epsilon_1 \geq \epsilon$, and $\delta \in [0, 1]$,
- sampling access to D.

provides sampling access to a distribution $\tilde{D}$ such that

(i) $d_{TV}(\tilde{D}, D) \leq \epsilon_1$;

(ii) $\tilde{D} \in \mathcal{P}$.

with probability $\geq 1 - \delta$ over the samples it draws and its internal randomness.

⋆ The query complexity: $q = q(\epsilon, \epsilon_1, \delta, \Omega)$.

- # samples from D it takes per query (to $\tilde{D}$) in the worst case.

Sampling Improver

- $\mathcal{P}$: a fixed and given distributions on Ω.
- A distribution D over $[n]$, $d_{TV}(D, \mathcal{P}) \leq \epsilon$.

An $(\epsilon, \epsilon_1, \epsilon_2)$-sampling improver for $\mathcal{P}$ is a randomized algorithm, which is given

- $\epsilon \in (0, 1]$, $\epsilon_1, \epsilon_2 \in [0, 1]$ s.t. $\epsilon_1 + \epsilon_2 \geq \epsilon$, and $\delta \in [0, 1]$
- $\texttt{ORACLE}_1$ access to D,

provides $\texttt{ORACLE}_2$ access to a distribution $\tilde{D}$ such that

(i) $d_{TV}(\tilde{D}, D) \leq \epsilon_1$;

(ii) $d_{TV}(\tilde{D}, \mathcal{P}) \leq \epsilon_2$.

with probability $\geq 1 - \delta$ over the answers from $\texttt{ORACLE}_1$ and its internal randomness.

⋆ The query complexity: $q = q(\epsilon, \epsilon_1, \epsilon_2, \delta, \Omega)$.

- # queries the algorithm makes to $\texttt{ORACLE}_1$ in the worst case.

Learning Algorithms (for a class of distributions $\mathcal{C}$)

An algorithm $\mathcal{L}$ which

- gets independent samples from an unknown distribution $D \in \mathcal{C}$
- has input $\epsilon > 0$;

output, with high probability, a hypothesis $\tilde{D}$ such that $d_{TV}(D, \tilde{D}) \leq \epsilon$.

- If $\tilde{D} \in \mathcal{C}$, then we said $\mathcal{L}$ is proper.

Connections to learning

From learning to correcting

Theorem 4.1

Let $\mathcal{C}$ be a class of distributions over Ω and $D \in \mathcal{C}$.

Suppose that there exists a learning algorithm $\mathcal{L}$ for $\mathcal{C}$ with sample complexity $q_{\mathcal{L}}$.

Then, for any property $\mathcal{P}$ of distributions, there exists a sampling corrector for $\mathcal{P}$ with sample complexity $q(\epsilon, \epsilon_1, \delta) = q_{\mathcal{L}}(\frac{\epsilon_1 - \epsilon}{2}, \delta)$.

- Run $\mathcal{L}$ on the unknown $D \in \mathcal{C}$ to learn (whp) hypothesis $\hat{D}$ such that $D_{TV}(D, \hat{D}) \leq \frac{\epsilon_1 - \epsilon}{2} \quad \Rightarrow \quad d_{TV}(\hat{D}, \mathcal{P}) \leq \frac{\epsilon_1 + \epsilon}{2}$.
- Find (e.g., exhaustive search) a distribution $\tilde{D} \in \mathcal{P}$ closest to $\hat{D}$, and use it to produce "corrected samples".

Example: correcting monotonicity

Monotone distributions

A distribution D is monotone if its probability mass function is non-increasing, that is, if $D(1) \geq D(2) \geq \ldots \geq D(n)$.

Birgé decomposition [Birgé 1987]

Given $\alpha > 0$, the corresponding Birgé-decomposition of $[n]$ is the partition

$$\mathcal{I}_\alpha = (I_1, I_2, \ldots, I_\ell),$$

where $\ell = \Theta\left(\frac{\ln(\alpha n+1)}{\alpha}\right) = \Theta\left(\frac{\log n}{\alpha}\right)$, $|I_k| = \lfloor (1+\alpha)^k \rfloor$, $1 \leq k \leq \ell$.

Flattened distribution

For a distribution D and parameter $\alpha > 0$,

$$\Phi_\alpha(D)(i) \triangleq D(I_k)/|I_k|,$$

for all $k \in [\ell]$ and $i \in I_k$.

Monotone distributions

A distribution D is monotone if its probability mass function is non-increasing, that is, if $D(1) \geq D(2) \geq \ldots \geq D(n)$.

Birgé decomposition [Birgé 1987]

Given $\alpha > 0$, the corresponding Birgé-decomposition of $[n]$ is the partition

$$\mathcal{I}_\alpha = (I_1, I_2, \ldots, I_\ell),$$

where $\ell = \Theta\left(\frac{\ln(\alpha n+1)}{\alpha}\right) = \Theta\left(\frac{\log n}{\alpha}\right)$, $|I_k| = \lfloor (1+\alpha)^k \rfloor$, $1 \leq k \leq \ell$.

Flattened distribution

For a distribution D and parameter $\alpha > 0$,

$$\Phi_\alpha(D)(i) \triangleq D(I_k)/|I_k|,$$

for all $k \in [\ell]$ and $i \in I_k$.

Monotone distributions

A distribution D is monotone if its probability mass function is non-increasing, that is, if $D(1) \geq D(2) \geq \ldots \geq D(n)$.

Birgé decomposition [Birgé 1987]

Given $\alpha > 0$, the corresponding Birgé-decomposition of $[n]$ is the partition

$$\mathcal{I}_\alpha = (I_1, I_2, \ldots, I_\ell),$$

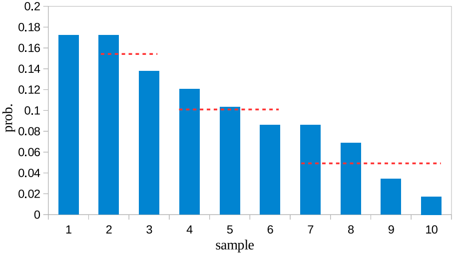
where $\ell = \Theta\left(\frac{\ln(\alpha n+1)}{\alpha}\right) = \Theta\left(\frac{\log n}{\alpha}\right)$, $|I_k| = \lfloor (1+\alpha)^k \rfloor$, $1 \leq k \leq \ell$.

Flattened distribution

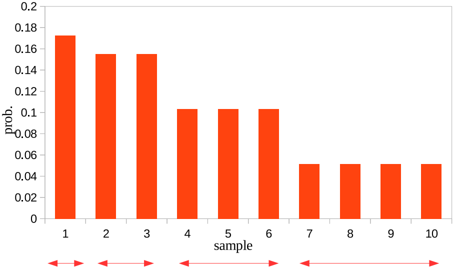
For a distribution D and parameter $\alpha > 0$,

$$\Phi_\alpha(D)(i) \triangleq D(I_k)/|I_k|,$$

for all $k \in [\ell]$ and $i \in I_k$.


prob.
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1 2 3 4 5 6 7 8 9 10
sample

D


prob.
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1 2 3 4 5 6 7 8 9 10
sample

$\Phi_\alpha(D)$

Sampling from $\Phi_\alpha(D)$

Sampling from $\Phi_\alpha(D)$ needs only one sample from D.

- We have the explicit Birgé decomposition $I_1, \ldots, I_\ell$ of $[n]$ at hand.
- Draw a sample x from D. Once you get it, find in which of these intervals it fell, say I_{49}. Forget now about x, and output a sample y drawn uniformly at random from I_{49}.
- **Claim:** y is exactly distributed according to $\Phi_\alpha(D)$.
  - For any given $i \in [\ell]$, we have that x belongs to I_i with prob. $D(I_i)$.
  - Conditioned on $i \in [\ell]$, y is uniformly distributed in I_i.
- We only need one sample from D to output a sample from $\Phi_\alpha(D)$ (along with some internal randomness for the second step).

Sampling from $\Phi_\alpha(D)$

Sampling from $\Phi_\alpha(D)$ needs only one sample from D.

- We have the explicit Birgé decomposition $I_1, \ldots, I_\ell$ of $[n]$ at hand.
- Draw a sample x from D. Once you get it, find in which of these intervals it fell, say I_{49}. Forget now about x, and output a sample y drawn uniformly at random from I_{49}.
- **Claim:** y is exactly distributed according to $\Phi_\alpha(D)$.
  - For any given $i \in [\ell]$, we have that x belongs to I_i with prob. $D(I_i)$.
  - Conditioned on $i \in [\ell]$, y is uniformly distributed in I_i.
- We only need one sample from D to output a sample from $\Phi_\alpha(D)$ (along with some internal randomness for the second step).

Sampling from $\Phi_\alpha(D)$

Sampling from $\Phi_\alpha(D)$ needs only one sample from D.

- We have the explicit Birgé decomposition $I_1, \ldots, I_\ell$ of $[n]$ at hand.
- Draw a sample x from D. Once you get it, find in which of these intervals it fell, say I_{49}. Forget now about x, and output a sample y drawn uniformly at random from I_{49}.
- **Claim:** y is exactly distributed according to $\Phi_\alpha(D)$.
  - For any given $i \in [\ell]$, we have that x belongs to I_i with prob. $D(I_i)$.
  - Conditioned on $i \in [\ell]$, y is uniformly distributed in I_i.
- We only need one sample from D to output a sample from $\Phi_\alpha(D)$ (along with some internal randomness for the second step).

Birgé flattening doesn't increase TV of two distributions

Claim 2.1

$d_{TV}(\Phi_\alpha(D_1), \Phi_\alpha(D_2)) \le d_{TV}(D_1, D_2)$.

- $$\begin{aligned} 2d_{TV}(\Phi_\alpha(D_1), \Phi_\alpha(D_2)) &= \sum_{i=1}^{n} |\Phi_\alpha(D_1)(i) - \Phi_\alpha(D_2)(i)| \\ &= \sum_{k=1}^{\ell} \sum_{i \in I_k} \left| \frac{D_1(I_k)}{|I_k|} - \frac{D_2(I_k)}{|I_k|} \right| \\ &= \sum_{k=1}^{\ell} |D_1(I_k) - D_2(I_k)| = \sum_{k=1}^{\ell} \left| \sum_{i \in I_k} (D_1(i) - D_2(i)) \right| \\ &\le \sum_{k=1}^{\ell} \sum_{i \in I_k} |D_1(i) - D_2(i)| \\ &= \sum_{i=1}^{n} |D_1(i) - D_2(i)| = 2d_{TV}(D_1, D_2). \end{aligned}$$

More facts on the flattened distribution

Theorem 2.4 [Birgé 1987]

If D is monotone, then $d_{TV}(D, \Phi_\alpha(D)) \leq \alpha$.

Corollary 2.5

Suppose D is ϵ-close to monotone, and $\alpha > 0$. Then,

- $d_{TV}(D, \Phi_\alpha(D)) \leq 2\epsilon + \alpha$.
- $\Phi_\alpha(D)$ is also ϵ-close to monotone.

- Let D' be a monotone distribution s.t. $d_{TV}(D, D') = \eta \leq \epsilon$.
- $d_{TV}(\Phi_\alpha(D) - \Phi_\alpha(D')) \leq d_{TV}(D, D') = \eta$ (Claim 2.1).
  - **Note:** $\Phi_\alpha(D')$ is monotone.
- $d_{TV}(D, \Phi_\alpha(D)) \leq d_{TV}(D, D') + d_{TV}(D', \Phi_\alpha(D')) + d_{TV}(\Phi_\alpha(D'), \Phi_\alpha(D))$.

More facts on the flattened distribution

Theorem 2.4 [Birgé 1987]

If D is monotone, then $d_{TV}(D, \Phi_\alpha(D)) \leq \alpha$.

Corollary 2.5

Suppose D is ϵ-close to monotone, and $\alpha > 0$. Then,

- $d_{TV}(D, \Phi_\alpha(D)) \leq 2\epsilon + \alpha$.
- $\Phi_\alpha(D)$ is also ϵ-close to monotone.

- Let D' be a monotone distribution s.t. $d_{TV}(D, D') = \eta \leq \epsilon$.
- $d_{TV}(\Phi_\alpha(D) - \Phi_\alpha(D')) \leq d_{TV}(D, D') = \eta$ (Claim 2.1).
  - **Note:** $\Phi_\alpha(D')$ is monotone.
- $d_{TV}(D, \Phi_\alpha(D)) \leq d_{TV}(D, D') + d_{TV}(D', \Phi_\alpha(D')) + d_{TV}(\Phi_\alpha(D'), \Phi_\alpha(D))$.

Correcting by learning

Lemma 5.1

Fix any constant $c > 0$. For any $\epsilon, \epsilon_1 \geq (3 + c)\epsilon$ and $\epsilon_2 = 0$, any type of oracle `ORACLE` and any number of queries m, there exists a sampling corrector for monotonicity from sampling to `ORACLE` with sample complexity $O(\log n/\epsilon^3)$.

- Learn a good approximation of the distribution to correct.
- Use this approximation to build a good monotone distribution offline (+searching via linear programming).

Sketch of the Proof of Lemma 5.1

- Consider the Birgé decomposition $\mathcal{I}_\alpha = (I_1, \ldots, I_\ell)$, $\alpha = c\epsilon/3$, $\ell = O(\log n/\epsilon)$.
- Learn, with $O(\frac{\log n}{\epsilon^3})$ samples, a flattened distribution $\bar{D}$, where $d_{TV}(D, \bar{D}) \leq 2\epsilon + \alpha$ (by [Birgé 1987] & Corollary 2.5).
  - Learn $\bar{D} \rightarrow$ getting $\bar{D}'$.
- $d_{TV}(\bar{D}, \mathcal{M}) = d_{TV}(\Phi_\alpha(D), \mathcal{M}) \leq d_{TV}(\Phi_\alpha(D), \Phi_\alpha(M)) \leq d_{TV}(D, M) \leq \epsilon$.
  - M: the closest monotone distribution to D.
  - ⋆ $\bar{D}'$ is $(\epsilon + \alpha)$-close to monotone.
- Find $M' \in \mathcal{M}$ closest to $\bar{D}'$ such that:

$$\text{minimize} \sum_{j=1}^{\ell} \left| x_j - \frac{\bar{D}'(I_j)}{|I_j|} \right| \cdot |I_j|$$

$$\text{subject to } 1 \geq x_1 \geq x_2 \geq \ldots \geq x_\ell \geq 0, \sum_{j=1}^{\ell} x_j |I_j| = 1.$$

  $M'(i) = x_{\text{ind}(i)}$, for $i \in I_{\text{ind}(i)}$.
- $d_{TV}(D, M') \leq d_{TV}(D, \bar{D}) + d_{TV}(\bar{D}, \bar{D}') + d_{TV}(\bar{D}', M') \leq 3\epsilon + 3\alpha = (3 + c)\epsilon$.

Sketch of the Proof of Lemma 5.1

- Consider the Birgé decomposition $\mathcal{I}_\alpha = (I_1, \ldots, I_\ell)$, $\alpha = c\epsilon/3$, $\ell = O(\log n/\epsilon)$.
- Learn, with $O(\frac{\log n}{\epsilon^3})$ samples, a flattened distribution $\bar{D}$, where $d_{TV}(D, \bar{D}) \leq 2\epsilon + \alpha$ (by [Birgé 1987] & Corollary 2.5).
  - Learn $\bar{D} \rightarrow$ getting $\bar{D}'$.
- $d_{TV}(\bar{D}, \mathcal{M}) = d_{TV}(\Phi_\alpha(D), \mathcal{M}) \leq d_{TV}(\Phi_\alpha(D), \Phi_\alpha(M)) \leq d_{TV}(D, M) \leq \epsilon$.
  - M: the closest monotone distribution to D.
  - ⋆ $\bar{D}'$ is $(\epsilon + \alpha)$-close to monotone.
- Find $M' \in \mathcal{M}$ closest to $\bar{D}'$ such that:

$$\text{minimize} \sum_{j=1}^{\ell} \left| x_j - \frac{\bar{D}'(I_j)}{|I_j|} \right| \cdot |I_j|$$

$$\text{subject to } 1 \geq x_1 \geq x_2 \geq \ldots \geq x_\ell \geq 0, \sum_{j=1}^{\ell} x_j |I_j| = 1.$$

$M'(i) = x_{\text{ind}(i)}$, for $i \in I_{\text{ind}(i)}$.

- $d_{TV}(D, M') \leq d_{TV}(D, \bar{D}) + d_{TV}(\bar{D}, \bar{D}') + d_{TV}(\bar{D}', M') \leq 3\epsilon + 3\alpha = (3 + c)\epsilon$.

Sketch of the Proof of Lemma 5.1

- Consider the Birgé decomposition $\mathcal{I}_\alpha = (I_1, \ldots, I_\ell)$, $\alpha = c\epsilon/3$, $\ell = O(\log n/\epsilon)$.
- Learn, with $O(\frac{\log n}{\epsilon^3})$ samples, a flattened distribution $\bar{D}$, where $d_{TV}(D, \bar{D}) \leq 2\epsilon + \alpha$ (by [Birgé 1987] & Corollary 2.5).
  - Learn $\bar{D} \rightarrow$ getting $\bar{D}'$.
- $d_{TV}(\bar{D}, \mathcal{M}) = d_{TV}(\Phi_\alpha(D), \mathcal{M}) \leq d_{TV}(\Phi_\alpha(D), \Phi_\alpha(M)) \leq d_{TV}(D, M) \leq \epsilon$.
  - M: the closest monotone distribution to D.
  - ⋆ $\bar{D}'$ is $(\epsilon + \alpha)$-close to monotone.
- Find $M' \in \mathcal{M}$ closest to $\bar{D}'$ such that:

$$\text{minimize} \sum_{j=1}^{\ell} \left| x_j - \frac{\bar{D}'(I_j)}{|I_j|} \right| \cdot |I_j|$$

$$\text{subject to } 1 \geq x_1 \geq x_2 \geq \ldots \geq x_\ell \geq 0, \sum_{j=1}^{\ell} x_j |I_j| = 1.$$

$M'(i) = x_{\text{ind}(i)}$, for $i \in I_{\text{ind}(i)}$.

- $d_{TV}(D, M') \leq d_{TV}(D, \bar{D}) + d_{TV}(\bar{D}, \bar{D}') + d_{TV}(\bar{D}', M') \leq 3\epsilon + 3\alpha = (3 + c)\epsilon$.

Sketch of the Proof of Lemma 5.1

- Consider the Birgé decomposition $\mathcal{I}_\alpha = (I_1, \ldots, I_\ell)$, $\alpha = c\epsilon/3$, $\ell = O(\log n/\epsilon)$.
- Learn, with $O(\frac{\log n}{\epsilon^3})$ samples, a flattened distribution $\bar{D}$, where $d_{TV}(D, \bar{D}) \leq 2\epsilon + \alpha$ (by [Birgé 1987] & Corollary 2.5).
  - Learn $\bar{D} \rightarrow$ getting $\bar{D}'$.
- $d_{TV}(\bar{D}, \mathcal{M}) = d_{TV}(\Phi_\alpha(D), \mathcal{M}) \leq d_{TV}(\Phi_\alpha(D), \Phi_\alpha(M)) \leq d_{TV}(D, M) \leq \epsilon$.
  - M: the closest monotone distribution to D.
  - ⋆ $\bar{D}'$ is $(\epsilon + \alpha)$-close to monotone.
- Find $M' \in \mathcal{M}$ closest to $\bar{D}'$ such that:

$$\text{minimize} \sum_{j=1}^{\ell} \left| x_j - \frac{\bar{D}'(I_j)}{|I_j|} \right| \cdot |I_j|$$

$$\text{subject to } 1 \geq x_1 \geq x_2 \geq \ldots \geq x_\ell \geq 0, \sum_{j=1}^{\ell} x_j |I_j| = 1.$$

  $M'(i) = x_{\text{ind}(i)}$, for $i \in I_{\text{ind}(i)}$.
- $d_{TV}(D, M') \leq d_{TV}(D, \bar{D}) + d_{TV}(\bar{D}, \bar{D}') + d_{TV}(\bar{D}', M') \leq 3\epsilon + 3\alpha = (3 + c)\epsilon$.

Sketch of the Proof of Lemma 5.1

- Consider the Birgé decomposition $\mathcal{I}_\alpha = (I_1, \ldots, I_\ell)$, $\alpha = c\epsilon/3$, $\ell = O(\log n/\epsilon)$.
- Learn, with $O(\frac{\log n}{\epsilon^3})$ samples, a flattened distribution $\bar{D}$, where $d_{TV}(D, \bar{D}) \leq 2\epsilon + \alpha$ (by [Birgé 1987] & Corollary 2.5).
  - Learn $\bar{D} \rightarrow$ getting $\bar{D}'$.
- $d_{TV}(\bar{D}, \mathcal{M}) = d_{TV}(\Phi_\alpha(D), \mathcal{M}) \leq d_{TV}(\Phi_\alpha(D), \Phi_\alpha(M)) \leq d_{TV}(D, M) \leq \epsilon$.
  - M: the closest monotone distribution to D.
  - ⋆ $\bar{D}'$ is $(\epsilon + \alpha)$-close to monotone.
- Find $M' \in \mathcal{M}$ closest to $\bar{D}'$ such that:

$$\text{minimize} \sum_{j=1}^{\ell} \left| x_j - \frac{\bar{D}'(I_j)}{|I_j|} \right| \cdot |I_j|$$

$$\text{subject to } 1 \geq x_1 \geq x_2 \geq \ldots \geq x_\ell \geq 0, \sum_{j=1}^{\ell} x_j |I_j| = 1.$$

$M'(i) = x_{\mathrm{ind}(i)}$, for $i \in I_{\mathrm{ind}(i)}$.

- $d_{TV}(D, M') \leq d_{TV}(D, \bar{D}) + d_{TV}(\bar{D}, \bar{D}') + d_{TV}(\bar{D}', M') \leq 3\epsilon + 3\alpha = (3 + c)\epsilon$.

Sketch of the Proof of Lemma 5.1

- Consider the Birgé decomposition $\mathcal{I}_\alpha = (I_1, \ldots, I_\ell)$, $\alpha = c\epsilon/3$, $\ell = O(\log n/\epsilon)$.
- Learn, with $O(\frac{\log n}{\epsilon^3})$ samples, a flattened distribution $\bar{D}$, where $d_{TV}(D, \bar{D}) \leq 2\epsilon + \alpha$ (by [Birgé 1987] & Corollary 2.5).
  - Learn $\bar{D} \rightarrow$ getting $\bar{D}'$.
- $d_{TV}(\bar{D}, \mathcal{M}) = d_{TV}(\Phi_\alpha(D), \mathcal{M}) \leq d_{TV}(\Phi_\alpha(D), \Phi_\alpha(M)) \leq d_{TV}(D, M) \leq \epsilon$.
  - M: the closest monotone distribution to D.
  - ⋆ $\bar{D}'$ is $(\epsilon + \alpha)$-close to monotone.
- Find $M' \in \mathcal{M}$ closest to $\bar{D}'$ such that:

$$\text{minimize} \sum_{j=1}^{\ell} \left| x_j - \frac{\bar{D}'(I_j)}{|I_j|} \right| \cdot |I_j|$$

$$\text{subject to } 1 \geq x_1 \geq x_2 \geq \ldots \geq x_\ell \geq 0, \sum_{j=1}^{\ell} x_j |I_j| = 1.$$

  $M'(i) = x_{\text{ind}(i)}$, for $i \in I_{\text{ind}(i)}$.
- $d_{TV}(D, M') \leq d_{TV}(D, \bar{D}) + d_{TV}(\bar{D}, \bar{D}') + d_{TV}(\bar{D}', M') \leq 3\epsilon + 3\alpha = (3 + c)\epsilon$.

Oblivious correcting of monotonicity

- Consider D which is $O(1/\log^2 n)$-close to monotone.

Corollary 5.5

For every $\epsilon' \in (0,1)$, there exists an (oblivious) sampling corrector for monotonicity of $O(1)$ sample complexity, with parameters $\epsilon = O(\epsilon'^3/\log^2 n)$, $\epsilon_1 = O(\epsilon')$.

High level idea:

- Treat D as a $O(\log n)$-histogram on the Birgé decomposition.
- *Implicitly* approximate it.
- Correct this histogram by adding a certain amount of prob. weight to every interval.

Lemma 5.2

- $\mathcal{I} = (I_1, \ldots, I_k)$: a Birgé decomposition of $[n]$, s.t. $|I_{j+1}|/|I_j| = 1 + c$ for all j.
- D: a k-histogram distribution on $\mathcal{I}$, ϵ-close to monotone.

Then, there exists a monotone distribution $\tilde{D}$, such that

- $\tilde{D}$ can be sampled in constant time from given oracle access to D;
- $d_{TV}(D, \tilde{D}) \leq O(\epsilon k^2)$.
- $\tilde{D}$ is also a k-histogram distribution on $\mathcal{I}$.

Claim 5.3

Let D be a k-histogram distribution on $\mathcal{I}$ that is ϵ-close to monotone. Then, for any $j \in [k-1]$,

$$D(I_{j+1}) \leq (1 + c) \cdot D(I_j) + \epsilon(2 + c).$$

Sketch of the proof of Lemma 5.2

- Claim 5.3 suggests a correcting scheme: output samples according to $\tilde{D}$, which is a k-histogram on $\mathcal{I}$ defined by

$$\begin{aligned}\tilde{D}(I_k) &= \lambda(D(I_k)) \\ \tilde{D}(I_{k-1}) &= \lambda(D(I_k) + \epsilon(2+c)) \\ &\vdots \\ \tilde{D}(I_j) &= \lambda\left(D(I_j) + (k-j)\epsilon(2+c)\right) \\ &= \lambda \cdot D(I_j) + (1-\lambda)\frac{k-j}{k(k-1)/2},\end{aligned}$$

$\lambda \triangleq \left(1 + \epsilon(2+c)\frac{k(k-1)}{2}\right)^{-1}$: normalizing factor.

- $$2d_{TV}(D, \tilde{D}) = \sum_{j=1}^{k} |D(I_j) - \tilde{D}(I_j)| \leq 1 - \frac{1 - \epsilon(2+c)\frac{k(k-1)}{2}}{1 + \epsilon(2+c)\frac{k(k-1)}{2}} = O(\epsilon k^2).$$

Correcting uniformity with scarce randomness

- Allowing arbitrary amounts of additional randomness makes the correcting task almost trivial.

⋆ Using roughly $\log|\Omega|$ random bits per query, then interpolate arbitrarily between D and the uniform distribution, say $\mathcal{U}$.

  - Sampling improver.

Von Neumann sampling corrector

Theorem 7.1

For any $\epsilon = \epsilon_1 < 0.49$, there exists a sampling corrector for $\mathcal{U}$ with query complexity $O(\log n \cdot (\log \log n + \log(1/\delta)))$, where δ is the failure probability per sample.

- Idea: see a draw from D as a biased coin toss.
  - Depending on whether the sample lands in $S_0 = \{1, \ldots, n/2\}$ or $S_1 = \{n/2+1, \ldots, n\}$.
  - ⋆ **Note:** $|D(S_0) - D(S_1)| \leq 2\epsilon$ by the assumption that $d_{TV}(D, \mathcal{U}) \leq \epsilon$.
  - Let $p \triangleq D(S_0)$, $p \in [1/2 - \epsilon, 1/2 + \epsilon]$.

Proof of Theorem 7.1 (contd.)

- Assume that $D_{TV}(D,\mathcal{U}) \le \epsilon < 1/2 - c$.
- Take at most $m = \left\lceil (\log^{-1} \frac{1}{1-c}) \log \frac{2}{\delta'} \right\rceil$ samples, and stop as soon as a sequence $S_0 S_1$ or $S_1 S_0$ is seen.
  - Output a bit 0 or 1 respectively.
  - If it does NOT happen, output FAIL.
  - ⋆ The probability of failure $\le p^m + (1-p)^m \le 2(1-c)^m \le \delta/(\log n)$.
- Extract $\log n$ random bits, output a uniform random number $s \in [n]$ w.p. $\ge 1 - \delta$.
  - Using $O(m \log n) = O(\log n \log \frac{\log n}{\delta})$ samples from D.
  - ⋆ Yet, $O(\log n)$ samples in expectation.

On convolutions of distributions over Abelian groups

Convolutions of distributions over a finite group

For any tow probability distributions D_1, D_2 over a finite group G, the *convolution* of D_1 and D_2 is defined by

$$D_1 * D_2(x) = \sum_{g \in G} D_1(xg^{-1}) D_2(g).$$

If G is Abelian, $D_1 * D_2 = D_2 * D_1$.

Fact [Maciej 2013]

Let G be a finite Abelian gruop, and D_1, D_2 be two probability distributions over G. Then,

$$d_{TV}(\mathcal{U}(G), D_1 * D_2) \leq 2 \cdot d_{TV}(\mathcal{U}(G), D_1) \cdot d_{TV}(\mathcal{U}(G), D_2).$$

Convolution improver

Theorem 7.2

For any $\epsilon < \frac{1}{2}$, ϵ_2, and $\epsilon_1 = \epsilon + \epsilon_2$, there exists a sampling improver for uniformity with query complexity $O(\frac{\log(1/\epsilon_2)}{\log(1/\epsilon)})$.

- **Idea:** drawing two independent samples $x, y \sim D$ and computing $z = (x + y \mod n) + 1$ guarantees that the distribution of z is $(2\epsilon^2)$-close to $\mathcal{U}$.
- Extending the above observation to a sum of $k := \frac{\log(1/\epsilon_2)}{\log(1/\epsilon)}$ independent elements $s_1, \ldots, s_k \sim D$ and computing $s = \left(\sum_{\ell=1}^{k} s_\ell \mod n\right) + 1 \in [n]$, the distribution $\tilde{D}$ of s is $((2\epsilon)^k/2)$-close to $\mathcal{U}$.
- Choose k such that $(2\epsilon)^k/2 = \epsilon_2$, we get $d_{TV}(D, \tilde{D}) \leq d_{TV}(D, \mathcal{U}) + d_{TV}(\mathcal{U}, \tilde{D}) \leq \epsilon + \epsilon_2$.

The problem of the convolution improver

- Say $D^{(k)} \triangleq D * \cdots * D$, $D^{(k)}$ could be a little bit far from D.
  - $d_{TV}(D, D^{(k)}) \leq \epsilon + 2^{k-1}\epsilon^k$.
- **Bad news:** There exists a distribution D on $\mathbb{Z}_n$ such that $d_{TV}(D,\mathcal{U}) = \epsilon$, yet $d_{TV}(D, D * D) = \epsilon + \frac{3}{4}\epsilon^2 + O(\epsilon^3)$.

Hybrid improver

Theorem 7.3

For any $\epsilon \leq \frac{1}{2}$, $\epsilon_1 = \frac{\epsilon}{2} + 2\epsilon^3 + \epsilon'$, and $\epsilon_2 = \frac{\epsilon}{2} + \epsilon'$, there exists a sampling improver for uniformity with query complexity $O(\frac{\log(1/\epsilon')}{\log(1/\epsilon)})$.

- **Idea:** $\tilde{D} \triangleq (1 - p_0)D + p_0 D^{(k)}$.
  - p_0: the probability that the two independently samples $s_1, s_2 \sim D$ located both in S_0 or both in S_1.
  - Getting a sample from $\tilde{D}$ only requires $\leq k + 2$ queries from D.

Theorem 7.4 (Bootstrapping)

For any $\epsilon \leq \frac{1}{2}$, $0 < \epsilon_2 < \epsilon$, and $\epsilon_1 = \epsilon - \epsilon_2 + O(\epsilon^3)$, there exists a sampling improver for uniformity with query complexity $O(\frac{\log^2(1/\epsilon_2)}{\log(1/\epsilon)})$.

Comparison with randomness extractors

- In the randomness extractor model:
  - One is provided with a source of *imperfect* random bits (sometimes an additional source of completely random bits).
  - **Goal:** output random bits (close to uniformly distributed) as many as possible.
- One could view extractors as sampling improvers for uniformity.
  - Both of them attempt to minimize the use of extra randomness.
- The differences:
  - Randomness extractors assume a lower bound on the min-entropy (i.e., $\log(1/\max_i D(i))$) of the input distribution, while sampling improvers assume the distribution to be ϵ-close to uniformity.
  - We have sampling improvers that do not use any extra random bits, not the case for randomized extractor constructions.
  - ...

Comparison with randomness extractors

- In the randomness extractor model:
  - One is provided with a source of *imperfect* random bits (sometimes an additional source of completely random bits).
  - **Goal:** output random bits (close to uniformly distributed) as many as possible.
- One could view extractors as sampling improvers for uniformity.
  - Both of them attempt to minimize the use of extra randomness.
- The differences:
  - Randomness extractors assume a lower bound on the min-entropy (i.e., $\log(1/\max_i D(i))$) of the input distribution, while sampling improvers assume the distribution to be ϵ-close to uniformity.
  - We have sampling improvers that do not use any extra random bits, not the case for randomized extractor constructions.
  - ...


Thank you.