DISTRIBUTED SIMULATION AND DISTRIBUTED INFERENCE

Last Thing Standing Between You and a Beer

Clément Canonne (Stanford University)

June 15, 2018

Joint work with Jayadev Acharya (Cornell University) and Himanshu Tyagi (IISc Bangalore)

A STORY

Boaty McBoatface is starting its first mission today!
It's going to Antarctica to study global warming, not to play.

The world's oceans are changing, you see.
It's freezing down there, but not as cold as it used to be.

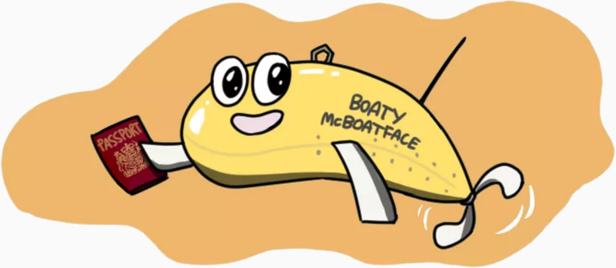


Illustration ©Dami Lee

Boaty's findings will be sent to scientists with care,
By way of a radio link, but with a certain flair.

Illustration ©Dami Lee

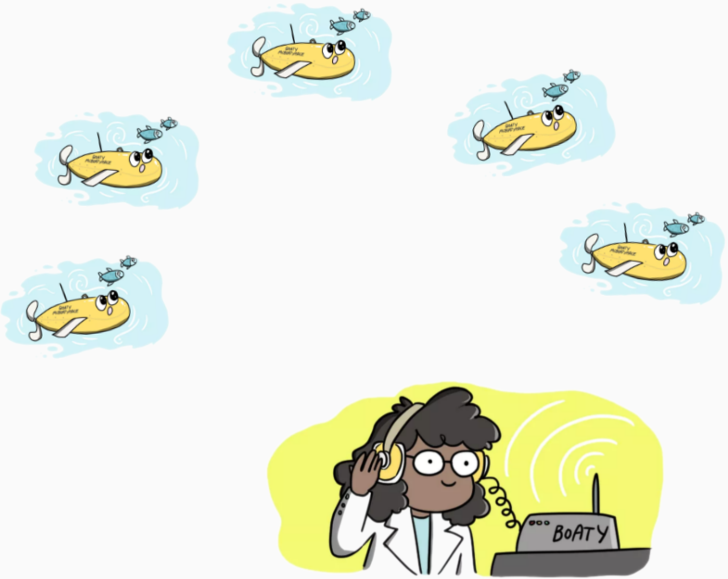


Illustration ©Dami Lee

McBoatfaces are expensive

What is the most ship-efficient protocol to reliably test whether the distribution of temperatures matches the one on record?

DISTRIBUTED INFERENCE

THE SETTING: "SIMULTANEOUS COMMUNICATION PROTOCOL" (SMP)

THE SETTING: "SIMULTANEOUS COMMUNICATION PROTOCOL" (SMP)

- an inference task $\mathcal{P}$ over k-ary distributions

THE SETTING: "SIMULTANEOUS COMMUNICATION PROTOCOL" (SMP)

- an inference task $\mathcal{P}$ over k-ary distributions
- an unknown k-ary distribution p

THE SETTING: "SIMULTANEOUS COMMUNICATION PROTOCOL" (SMP)

- an inference task $\mathcal{P}$ over k-ary distributions
- an unknown k-ary distribution p
- one centralized "referee" $\mathcal{R}$ who needs to solve $\mathcal{P}$ on p

THE SETTING: "SIMULTANEOUS COMMUNICATION PROTOCOL" (SMP)

- an inference task $\mathcal{P}$ over k-ary distributions
- an unknown k-ary distribution p
- one centralized "referee" $\mathcal{R}$ who needs to solve $\mathcal{P}$ on p
- n communication-limited players, each can send ℓ bits to $\mathcal{R}$

THE SETTING: "SIMULTANEOUS COMMUNICATION PROTOCOL" (SMP)

- an inference task $\mathcal{P}$ over k-ary distributions
- an unknown k-ary distribution p
- one centralized "referee" $\mathcal{R}$ who needs to solve $\mathcal{P}$ on p
- n communication-limited players, each can send ℓ bits to $\mathcal{R}$
- each player independently gets one sample from p

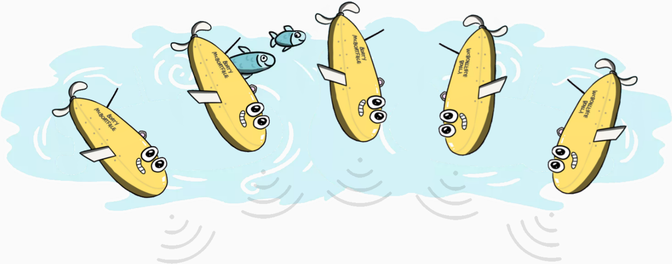
THE SETTING: "SIMULTANEOUS COMMUNICATION PROTOCOL" (SMP)

- an inference task $\mathcal{P}$ over k-ary distributions
- an unknown k-ary distribution p
- one centralized "referee" $\mathcal{R}$ who needs to solve $\mathcal{P}$ on p
- n communication-limited players, each can send ℓ bits to $\mathcal{R}$
- each player independently gets one sample from p

Question

As a function of k, ℓ, and all relevant parameters of $\mathcal{P}$, how many players n are required?

THE SETTING, CONT'D



THE SETTING, CONT'D

- Can assume $\ell < \log_2 k$, otherwise trivial

- Can assume $\ell < \log_2 k$, otherwise trivial
- Inference tasks: density estimation, parameter estimation, functional estimation, hypothesis testing/property testing...

THE SETTING, CONT'D

- Can assume $\ell < \log_2 k$, otherwise trivial
- Inference tasks: density estimation, parameter estimation, functional estimation, hypothesis testing/property testing...
- Different flavors: public-coin, pairwise-coin, private-coin

“SIMULATE-AND-INFER”

ONE APPROACH TO SOLVE IT ALL

Key Observation

If the referee can simulate independent samples from p using the messages from the players, then it can do anything.

Key Observation

If the referee can simulate independent samples from p using the messages from the players, then it can do anything.

Begging the question

Can the referee simulate independent samples from p using the messages from the players?

NO APPROACH TO SOLVE IT ALL?

Theorem

For every $k \geq 1$ and $\ell < \log k$, there exists no SMP with ℓ bits of communication per player for distributed simulation over $[k]$ with any finite number of players. (Even allowing public-coin and interactive protocols.)

Theorem

For every $k \geq 1$ and $\ell < \log k$, there exists no SMP with ℓ bits of communication per player for distributed simulation over $[k]$ with any finite number of players. (Even allowing public-coin and interactive protocols.)

Proof.

By contradiction, [...] pigeonhole principle [...]. $\square$

ONE APPROACH TO SOLVE IT ALL!

ONE APPROACH TO SOLVE IT ALL!

Theorem

For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$, with expected number of players $O(k/2^\ell \vee 1)$. Moreover, this is optimal even allowing public-coin and interactive protocols.

Theorem

For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$, with expected number of players $O(k/2^\ell \vee 1)$. Moreover, this is optimal even allowing public-coin and interactive protocols.

Proof.

Case $\ell = 1$.

Theorem

For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$, with expected number of players $O(k/2^\ell \vee 1)$. Moreover, this is optimal even allowing public-coin and interactive protocols.

Proof.

Case $\ell = 1$. Player $2i - 1$ and $2i$ both send 1 if their sample "hits" i;

Theorem

For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$, with expected number of players $O(k/2^\ell \vee 1)$. Moreover, this is optimal even allowing public-coin and interactive protocols.

Proof.

Case $\ell = 1$. Player $2i - 1$ and $2i$ both send 1 if their sample "hits" i; the referee outputs i if (i) player $2i - 1$ is the only odd player sending 1, and player $2i$ sends 0.

Theorem

For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$, with expected number of players $O(k/2^\ell \vee 1)$. Moreover, this is optimal even allowing public-coin and interactive protocols.

Proof.

Case $\ell = 1$. Player $2i - 1$ and $2i$ both send 1 if their sample "hits" i; the referee outputs i if (i) player $2i - 1$ is the only odd player sending 1, and player $2i$ sends 0. Then, conditioned on $\mathcal{R}$ not outputting $\perp$, i is outputted with probability p_i.

Theorem

For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$, with expected number of players $O(k/2^\ell \vee 1)$. Moreover, this is optimal even allowing public-coin and interactive protocols.

Proof.

Case $\ell = 1$. Player $2i - 1$ and $2i$ both send 1 if their sample "hits" i; the referee outputs i if (i) player $2i - 1$ is the only odd player sending 1, and player $2i$ sends 0. Then, conditioned on $\mathcal{R}$ not outputting $\perp$, i is outputted with probability p_i. And the probability to output $\perp$ is

$$1 - \prod_{i=1}^{k}(1 - p_i) \leq 1 - \phi(\|p\|_2)$$

Theorem

For every $k, \ell \geq 1$, there exists a private-coin protocol with ℓ bits of communication per player for distributed simulation over $[k]$, with expected number of players $O(k/2^\ell \vee 1)$. Moreover, this is optimal even allowing public-coin and interactive protocols.

Proof.

Case $\ell = 1$. Player $2i - 1$ and $2i$ both send 1 if their sample "hits" i; the referee outputs i if (i) player $2i - 1$ is the only odd player sending 1, and player 2i sends 0. Then, conditioned on $\mathcal{R}$ not outputting $\perp$, i is outputted with probability p_i. And the probability to output $\perp$ is

$$1 - \prod_{i=1}^{k}(1 - p_i) \leq 1 - \phi(\|p\|_2)$$

(and some complications to bound this away from 1). □

ONE APPROACH TO SOLVE IT ALL!

Corollary (Informal)

For any inference task $\mathcal{P}$ over k-ary distributions with sample complexity s in the non-distributed model, there is a private-coin protocol for $\mathcal{P}$, with ℓ bits of communication per player, and $n = O(s \cdot k/2^{\ell})$ players.

Illustration ©Dami Lee

ONE APPROACH TO SOLVE IT ALL!

Corollary (Learning in Total Variation)

For every $k, \ell \leq \log_2 k$, there is a private-coin protocol for learning k-ary distributions with ℓ bits per player, and $n = O(\frac{k^2}{2^\ell \varepsilon^2})$ players. (And this is optimal, even for public-coin and interactive protocols.)

Corollary (Learning in Total Variation)

For every $k, \ell \leq \log_2 k$, there is a private-coin protocol for learning k-ary distributions with ℓ bits per player, and $n = O(\frac{k^2}{2^\ell \varepsilon^2})$ players. (And this is optimal, even for public-coin and interactive protocols.)

Corollary (Testing Uniformity)

For every $k, \ell \leq \log_2 k$, there is a private-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O(\frac{k^{3/2}}{2^\ell \varepsilon^2})$ players.

ONE APPROACH TO REALLY, REALLY SOLVE IT ALL?

Natural Question

Is this “simulate-and-infer” approach optimal?

Natural Question

Is this “simulate-and-infer” approach optimal?

Conjecture (The Flying Pony Question)

Does the simulate-and-infer scheme that simulates independent samples *compressed to the size** of the problem using private-coin protocols, and sends them to the referee who then infers from them, always require the lowest number of players?

NO FLYING PONY

The answer is no:

Theorem

There exists an inference task $\mathcal{P}$ over k-ary distributions with $2^{\mathsf{size}(\mathcal{P})} \cdot \mathsf{samplecomplexity}(\mathcal{P}) = \Omega(k^{3/2})$, yet for which there is a 1-bit private-coin protocol with $n = O(k)$ players.

The answer is no:

Theorem

There exists an inference task $\mathcal{P}$ over k-ary distributions with $2^{\text{size}(\mathcal{P})} \cdot \text{samplecomplexity}(\mathcal{P}) = \Omega(k^{3/2})$, yet for which there is a 1-bit private-coin protocol with $n = O(k)$ players.

Proof.

Promise problem: p is either uniform, or uniform on an arbitrary subset of $k/2$ elements.

The answer is no:

Theorem

There exists an inference task $\mathcal{P}$ over k-ary distributions with $2^{\mathsf{size}(\mathcal{P})} \cdot \mathsf{samplecomplexity}(\mathcal{P}) = \Omega(k^{3/2})$, yet for which there is a 1-bit private-coin protocol with $n = O(k)$ players.

Proof.

Promise problem: p is either uniform, or uniform on an arbitrary subset of $k/2$ elements. $\mathsf{samplecomplexity}(\mathcal{P}) = \sqrt{k}$ (folklore); $2^{\mathsf{size}(\mathcal{P})} = \Omega(k)$ (from other theorems);

The answer is no:

Theorem

There exists an inference task $\mathcal{P}$ over k-ary distributions with $2^{\mathsf{size}(\mathcal{P})} \cdot \mathsf{samplecomplexity}(\mathcal{P}) = \Omega(k^{3/2})$, yet for which there is a 1-bit private-coin protocol with $n = O(k)$ players.

Proof.

Promise problem: p is either uniform, or uniform on an arbitrary subset of $k/2$ elements. $\mathsf{samplecomplexity}(\mathcal{P}) = \sqrt{k}$ (folklore); $2^{\mathsf{size}(\mathcal{P})} = \Omega(k)$ (from other theorems); very simple scheme with $O(k)$ players...

The answer is no:

Theorem

There exists an inference task $\mathcal{P}$ over k-ary distributions with $2^{\text{size}(\mathcal{P})} \cdot \textsf{samplecomplexity}(\mathcal{P}) = \Omega(k^{3/2})$, yet for which there is a 1-bit private-coin protocol with $n = O(k)$ players.

Proof.

Promise problem: p is either uniform, or uniform on an arbitrary subset of $k/2$ elements. $\textsf{samplecomplexity}(\mathcal{P}) = \sqrt{k}$ (folklore); $2^{\text{size}(\mathcal{P})} = \Omega(k)$ (from other theorems); very simple scheme with $O(k)$ players... everyone focuses on the first element. □

PUBLIC-COIN UNIFORMITY TESTING

Must decide:

$$p = u_k \text{ (uniform)}$$

Must decide:

$$p = u_k \text{ (uniform), or } \ell_1(p, u_k) > \varepsilon?$$

UNIFORMITY TESTING, RECAP

Must decide:

$$p = u_k \text{ (uniform), or } \ell_1(p, u_k) > \varepsilon?$$

(and be correct on any p with probability at least 2/3)

Fundamental property of distributions, building block for testing many others. [BKR04, Gol16, CDGR17]

UNIFORMITY TESTING, RECAP

Must decide:

$$p = u_k \text{ (uniform), or } \ell_1(p, u_k) > \varepsilon?$$

(and be correct on any p with probability at least 2/3)

Fundamental property of distributions, building block for testing many others. [BKR04, Gol16, CDGR17]

- completely understood in the non-distributed setting: $n = \Theta(\sqrt{k}/\varepsilon^2)$ samples [GR00, BFR+00, Pan08, DGPP17]

UNIFORMITY TESTING, RECAP

Must decide:

$$p = u_k \text{ (uniform), or } \ell_1(p, u_k) > \varepsilon?$$

(and be correct on any p with probability at least 2/3)

Fundamental property of distributions, building block for testing many others. [BKR04, Gol16, CDGR17]

- completely understood in the non-distributed setting: $n = \Theta(\sqrt{k}/\varepsilon^2)$ samples [GR00, BFR+00, Pan08, DGPP17]
- general "simulate-and-infer" scheme gives private-coin protocol with $n = O(k^{3/2}/\varepsilon^2)$ players

Must decide:

$$p = u_k \text{ (uniform), or } \ell_1(p, u_k) > \varepsilon?$$

(and be correct on any p with probability at least 2/3)

Fundamental property of distributions, building block for testing many others. [BKR04, Gol16, CDGR17]

- completely understood in the non-distributed setting: $n = \Theta(\sqrt{k}/\varepsilon^2)$ samples [GR00, BFR+00, Pan08, DGPP17]
- general "simulate-and-infer" scheme gives private-coin protocol with $n = O(k^{3/2}/\varepsilon^2)$ players (optimal?)

UNIFORMITY TESTING, RECAP

Must decide:

$$p = u_k \text{ (uniform), or } \ell_1(p, u_k) > \varepsilon?$$

(and be correct on any p with probability at least 2/3)

Fundamental property of distributions, building block for testing many others. [BKR04, Gol16, CDGR17]

- completely understood in the non-distributed setting: $n = \Theta(\sqrt{k}/\varepsilon^2)$ samples [GR00, BFR$^+$00, Pan08, DGPP17]
- general "simulate-and-infer" scheme gives private-coin protocol with $n = O(k^{3/2}/\varepsilon^2)$ players (optimal?)
- what if we allow public coins?

Theorem (Upper Bound)

For every $k, \ell \leq \log_2 k$, there is a public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ players.

Theorem (Upper Bound)

For every $k, \ell \leq \log_2 k$, there is a public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ players.

Theorem (Lower Bound)

This is optimal.

Theorem (Warm Up)

For every k, there is a public-coin protocol for testing uniformity over $[k]$ with $\ell = 1$ bit per player, and $n = O\left(\frac{k}{\varepsilon^3} \log \frac{1}{\varepsilon}\right)$ players.

Theorem (Warm Up)

For every k, there is a public-coin protocol for testing uniformity over $[k]$ with $\ell = 1$ bit per player, and $n = O\left(\frac{k}{\varepsilon^3} \log \frac{1}{\varepsilon}\right)$ players.

Proof.

Starting point: if p is ε-far from uniform, by definition,

$$\mathbb{E}_{x \sim u}[|p(x) - 1/k|] > \varepsilon/k\,.$$

Theorem (Warm Up)

For every k, there is a public-coin protocol for testing uniformity over $[k]$ with $\ell = 1$ bit per player, and $n = O\left(\frac{k}{\varepsilon^3} \log \frac{1}{\varepsilon}\right)$ players.

Proof.

Starting point: if p is ε-far from uniform, by definition,

$$\mathbb{E}_{x \sim u}[|p(x) - 1/k|] > \varepsilon/k\,.$$

Now, by an averaging argument (Markov),

$$\Pr_{x \sim u}[p(x) < (1 - \varepsilon/2)/k] > \varepsilon/2$$

Theorem (Warm Up)

For every k, there is a public-coin protocol for testing uniformity over [k] with $\ell = 1$ bit per player, and $n = O\left(\frac{k}{\varepsilon^3} \log \frac{1}{\varepsilon}\right)$ players.

Proof.

Starting point: if p is ε-far from uniform, by definition,

$$\mathbb{E}_{x\sim u}[|p(x) - 1/k|] > \varepsilon/k\,.$$

Now, by an averaging argument (Markov),

$$\Pr_{x\sim u}[p(x) < (1-\varepsilon/2)/k] > \varepsilon/2$$

and therefore [...] □

Theorem

For every k, there is a public-coin protocol for testing uniformity over $[k]$ with $\ell = 1$ bit per player, and $n = O(k/\varepsilon^2)$ players.

Theorem

For every k, there is a public-coin protocol for testing uniformity over $[k]$ with $\ell = 1$ bit per player, and $n = O(k/\varepsilon^2)$ players.

Proof.

Same starting point. Now, by a better averaging argument (Levin's work investment strategy), there exists $1 \leq j \leq L := \log_2(1/\varepsilon)$ s.t.

$$\Pr_{x \sim u}[p(x) < (1 - 2^{-j})/k] > \varepsilon \cdot 2^j/(L + 1 - j)^2$$

Theorem

For every k, there is a public-coin protocol for testing uniformity over $[k]$ with $\ell = 1$ bit per player, and $n = O(k/\varepsilon^2)$ players.

Proof.

Same starting point. Now, by a better averaging argument (Levin's work investment strategy), there exists $1 \leq j \leq L := \log_2(1/\varepsilon)$ s.t.

$$\Pr_{x\sim u}[p(x) < (1 - 2^{-j})/k] > \varepsilon \cdot 2^j/(L + 1 - j)^2$$

and therefore [...] (also, don't pay for the union bound!) □

Theorem

For every $k, \ell \leq \log_2 k$, there is a public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O(k/(2^{\ell/2}\varepsilon^2))$ players.

Theorem

For every $k, \ell \leq \log_2 k$, there is a public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O(k/(2^{\ell/2}\varepsilon^2))$ players.

Proof.

Starting point: for a set $S \subseteq [k]$ of $2^\ell - 1$ elements with $p(S) \simeq 2^\ell/k$, testing uniformity of the conditional distribution p_S would cost

$$(k/2^\ell) \cdot \sqrt{2^\ell}/\varepsilon^2 = k/(2^{\ell/2}\varepsilon^2)$$

samples, by rejection sampling.

Theorem

For every $k, \ell \leq \log_2 k$, there is a public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O(k/(2^{\ell/2}\varepsilon^2))$ players.

Proof.

Starting point: for a set $S \subseteq [k]$ of $2^\ell - 1$ elements with $p(S) \simeq 2^\ell/k$, testing uniformity of the conditional distribution p_S would cost

$$(k/2^\ell) \cdot \sqrt{2^\ell}/\varepsilon^2 = k/(2^{\ell/2}\varepsilon^2)$$

samples, by rejection sampling. Now, if p is ε-far from uniform then, on a u.a.r. set $S \subseteq [k]$ of $2^\ell - 1$ elements, p_S is ε-far from uniform on expectation.

Theorem

For every $k, \ell \leq \log_2 k$, there is a public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O(k/(2^{\ell/2}\varepsilon^2))$ players.

Proof.

Starting point: for a set $S \subseteq [k]$ of $2^\ell - 1$ elements with $p(S) \simeq 2^\ell/k$, testing uniformity of the conditional distribution p_S would cost

$$(k/2^\ell) \cdot \sqrt{2^\ell}/\varepsilon^2 = k/(2^{\ell/2}\varepsilon^2)$$

samples, by rejection sampling. Now, if p is ε-far from uniform then, on a u.a.r. set $S \subseteq [k]$ of $2^\ell - 1$ elements, p_S is ε-far from uniform on expectation.

Theorem

For every $k, \ell \leq \log_2 k$, there is a public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, and $n = O(k/(2^{\ell/2}\varepsilon^2))$ players.

Proof.

Starting point: for a set $S \subseteq [k]$ of $2^\ell - 1$ elements with $p(S) \simeq 2^\ell/k$, testing uniformity of the conditional distribution p_S would cost

$$(k/2^\ell) \cdot \sqrt{2^\ell}/\varepsilon^2 = k/(2^{\ell/2}\varepsilon^2)$$

samples, by rejection sampling. Now, if p is ε-far from uniform then, on a u.a.r. set $S \subseteq [k]$ of $2^\ell - 1$ elements, p_S is ε-far from uniform on expectation. Then, same ideas as before: Levin's strategy+careful allocation of the failure probabilities. □

Theorem

For every $k, \ell \leq \log_2 k$, every public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, must have $n = \Omega(k/(2^{\ell/2}\varepsilon^2))$ players.

Theorem

For every $k, \ell \leq \log_2 k$, every public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, must have $n = \Omega(k/(2^{\ell/2}\varepsilon^2))$ players.

Proof.

By Le Cam's two-point method, consider a distribution over "hard instances":

$$\forall 1 \leq i \leq k/2, \qquad p(2i-1), p(2i) = \left(\frac{1 \pm \varepsilon}{k}, \frac{1 \mp \varepsilon}{k}\right)$$

uniformly and independently at random. (Paninski's construction [Pan08]).

Theorem

For every $k, \ell \leq \log_2 k$, every public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, must have $n = \Omega(k/(2^{\ell/2}\varepsilon^2))$ players.

Proof.

By Le Cam's two-point method, consider a distribution over "hard instances":

$$\forall 1 \leq i \leq k/2, \qquad p(2i-1), p(2i) = \left(\frac{1 \pm \varepsilon}{k}, \frac{1 \mp \varepsilon}{k}\right)$$

uniformly and independently at random. (Paninski's construction [Pan08]). But needs to upper bound the TV distance between (i) distribution of n messages sent to the referee when $p = u_k$, and (ii) distribution of n messages under average hard instance.

Theorem

For every $k, \ell \leq \log_2 k$, every public-coin protocol for testing uniformity over $[k]$ with ℓ bits per player, must have $n = \Omega(k/(2^{\ell/2}\varepsilon^2))$ players.

Proof.

By Le Cam's two-point method, consider a distribution over "hard instances":

$$\forall 1 \leq i \leq k/2, \qquad p(2i-1), p(2i) = \left(\frac{1 \pm \varepsilon}{k}, \frac{1 \mp \varepsilon}{k}\right)$$

uniformly and independently at random. (Paninski's construction [Pan08]). But needs to upper bound the TV distance between (i) distribution of n messages sent to the referee when $p = u_k$, and (ii) distribution of n messages under average hard instance. The latter is not a product distribution… □

CONCLUSION

- General framework for distributed inference problems over discrete distributions, in the communication-starved regime

CONCLUSION

- General framework for distributed inference problems over discrete distributions, in the communication-starved regime
- Tight bounds for distributed simulation (and distributed learning [DGL+17, HMÖW18, HÖW18])

- General framework for distributed inference problems over discrete distributions, in the communication-starved regime
- Tight bounds for distributed simulation (and distributed learning [DGL+17, HMÖW18, HÖW18])
- First work on distributed testing

- General framework for distributed inference problems over discrete distributions, in the communication-starved regime
- Tight bounds for distributed simulation (and distributed learning [DGL+17, HMÖW18, HÖW18])
- First work on distributed testing
- Optimal protocols for public-coin uniformity testing

CONCLUSION

- General framework for distributed inference problems over discrete distributions, in the communication-starved regime
- Tight bounds for distributed simulation (and distributed learning [DGL+17, HMÖW18, HÖW18])
- First work on distributed testing
- Optimal protocols for public-coin uniformity testing
- Many questions and directions to explore*

Thank you

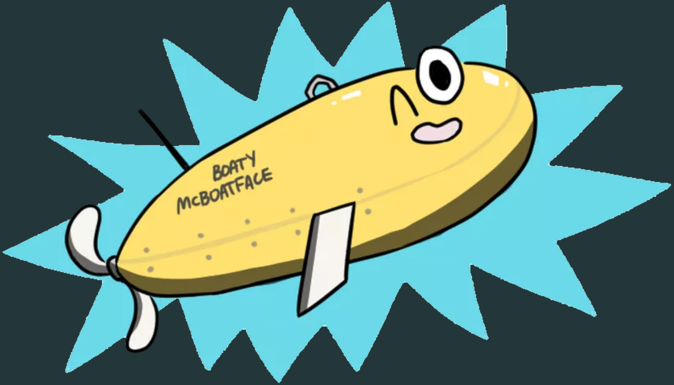


Illustration ©Dami Lee

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White.
Testing that distributions are close.
In Proceedings of FOCS, pages 189–197, 2000.

Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld.
Sublinear algorithms for testing monotone and unimodal distributions.
In Proceedings of STOC, pages 381–390, New York, NY, USA, 2004. ACM.

Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld.
Testing shape restrictions of discrete distributions.
Theory of Computing Systems, pages 1–59, 2017.

Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt.
Communication-efficient distributed learning of discrete distributions.
In Proceedings of NIPS, pages 6394–6404, 2017.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price.
Sample-optimal identity testing with high probability.
Electronic Colloquium on Computational Complexity (ECCC), 24:133, 2017.

Oded Goldreich.
The uniform distribution is complete with respect to testing identity to a fixed distribution.
Electronic Colloquium on Computational Complexity (ECCC), 23:15, 2016.

Oded Goldreich and Dana Ron.
On testing expansion in bounded-degree graphs.
Technical Report TR00-020, Electronic Colloquium on Computational Complexity (ECCC), 2000.

Yanjun Han, Pritam Mukherjee, Ayfer Özgür, and Tsachy Weissman.

Distributed statistical estimation of high-dimensional and nonparametric distributions with communication constraints, February 2018.
Talk given at ITA 2018.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman.
Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints.
ArXiv e-prints, February 2018.
abs/1802.08417.

Liam Paninski.
A coincidence-based test for uniformity given very sparsely sampled discrete data.
IEEE Transactions on Information Theory, 54(10):4750–4755, 2008.