Complexity of distributions and average-case hardness

Authors:

Dmitry Itsykson, Alexander Knop, Dmitry Sokolov

Institute:

St. Petersburg Department of V.A. Steklov Institute of Mathematics of the Russian Academy of Sciences

Definitions

SAMPLABLE DISTRIBUTIONS

Ensemble of distributions $D \in \mathbf{Samp}(n^k)$ iff there is a randomized $O(n^k)$-time algorithm A such that D_n and $A(1^n)$ are equally distributed.
We also denote $\mathbf{PSamp} = \bigcup_k \mathbf{Samp}(n^k)$.

HEURISTIC COMPUTATIONS

Distributional problem $(L, D) \in \mathrm{Heur}_\delta\mathbf{DTime}(n^k)$ iff there is $O(n^k)$-time algorithm A such that

$$\forall n \in \mathbb{N} \Pr_{x \leftarrow D_n}[A(x) = L(x)] > 1 - \delta.$$

Additionally denote $\mathrm{Heur}_\delta\mathbf{P} = \bigcup_k \mathrm{Heur}_\delta\mathbf{DTime}(n^k)$.

Definitions

SAMPLABLE DISTRIBUTIONS

Ensemble of distributions $D \in \mathbf{Samp}(n^k)$ iff there is a randomized $O(n^k)$-time algorithm A such that D_n and $A(1^n)$ are equally distributed.
We also denote $\mathbf{PSamp} = \bigcup_k \mathbf{Samp}(n^k)$.

HEURISTIC COMPUTATIONS

Distributional problem $(L, D) \in \mathrm{Heur}_\delta\mathbf{DTime}(n^k)$ iff there is $O(n^k)$-time algorithm A such that

$$\forall n \in \mathbb{N} \Pr_{x \leftarrow D_n}[A(x) = L(x)] > 1 - \delta.$$

Additionally denote $\mathrm{Heur}_\delta\mathbf{P} = \bigcup_k \mathrm{Heur}_\delta\mathbf{DTime}(n^k)$.

Hamiltonian Path

GUREVICH AND SHELAH, 1987

Let HP denote the language of Hamiltonian graphs. Then $(\mathrm{HP}, U) \in \mathrm{Heur}_{\frac{1}{2^{O(\sqrt{n})}}}\mathbf{DTime}(n)$.

OPEN PROBLEM

Find polynomial-time algorithm for HP.

Hamiltonian Path

GUREVICH AND SHELAH, 1987

Let HP denote the language of Hamiltonian graphs. Then $(\mathrm{HP}, U) \in \mathrm{Heur}_{\frac{1}{2^{O(\sqrt{n})}}}\mathbf{DTime}(n)$.

OPEN PROBLEM

Find polynomial-time algorithm for HP.

Graph Isomorphism

BABAI, ERDOS AND SELKOW, 1980

Let GI denote the language of pairs of isomorphic graphs. Then $(\mathrm{GI}, U) \in \mathrm{Heur}_{\frac{1}{\sqrt[7]{n}}}\mathbf{DTime}(n)$.

OPEN PROBLEM

Find polynomial-time algorithm for GI.

Graph Isomorphism

BABAI, ERDOS AND SELKOW, 1980

Let GI denote the language of pairs of isomorphic graphs. Then $(\mathrm{GI}, U) \in \mathrm{Heur}_{\frac{1}{\sqrt[7]{n}}}\mathbf{DTime}(n)$.

OPEN PROBLEM

Find polynomial-time algorithm for GI.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_\delta\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Goal and Result

GOAL

For every k there are a language L, ensemble D and small δ such that

1. $D \in \mathbf{PSamp}$;
2. $(L, D) \notin \mathrm{Heur}_{1-\delta}\mathbf{P}$;
3. for every $R \in \mathbf{Samp}(n^k)$ we have that $(L, R) \in \mathrm{Heur}_{\delta}\mathbf{DTime}(n)$.

RESULT

There are a language L and ensemble D such that

1. $D \in \mathbf{Samp}(n^{\log n})$;
2. $(L, D) \notin \mathrm{Heur}_{1-\frac{1}{2^{\log\log\log n}}}\mathbf{P}$;
3. for every $R \in \mathbf{PSamp}$ we have that $(L, R) \in \mathrm{Heur}_{\frac{1}{2^{\log\log\log n}}}\mathbf{DTime}(n)$.

Equivalent reformulations

DISTRIBUTIONAL PROBLEMS

Functions f and g satisfy CD property with parameters $\alpha(n) > 0$ and $\beta(n) > 0$ ($\mathrm{CD}_{\alpha(n),\beta(n)}(f(n), g(n))$) if there are $D \in \textbf{Samp}(f(n))$ and L such that

1. $(L, F) \in \mathrm{Heur}_{\alpha(n)}\mathbf{P}$ for every $F \in \textbf{Samp}(g(n))$.
2. $(L, D) \notin \mathrm{Heur}_{1-\beta(n)}\mathbf{P}$.

SAMPLING DISTRIBUTIONS

Functions f and g satisfy SD property with parameter $\lambda(n)$ ($\mathrm{SD}_{\lambda(n)}(f(n), g(n))$) if there is $D \in \textbf{Samp}(f(n))$ such that for every $F \in \textbf{Samp}(g(n))$, for infinitely many n the statistical distance between D_n and F_n is at least $1-\lambda(n)$.

Equivalent reformulations

DISTRIBUTIONAL PROBLEMS

Functions f and g satisfy CD property with parameters $\alpha(n) > 0$ and $\beta(n) > 0$ ($\mathrm{CD}_{\alpha(n),\beta(n)}(f(n), g(n))$) if there are $D \in \mathbf{Samp}(f(n))$ and L such that

1. $(L, F) \in \mathrm{Heur}_{\alpha(n)}\mathbf{P}$ for every $F \in \mathbf{Samp}(g(n))$.
2. $(L, D) \notin \mathrm{Heur}_{1-\beta(n)}\mathbf{P}$.

SAMPLING DISTRIBUTIONS

Functions f and g satisfy SD property with parameter $\lambda(n)$ ($\mathrm{SD}_{\lambda(n)}(f(n), g(n))$) if there is $D \in \mathbf{Samp}(f(n))$ such that for every $F \in \mathbf{Samp}(g(n))$, for infinitely many n the statistical distance between D_n and F_n is at least $1-\lambda(n)$.

Equivalent reformulations

$1 \to 2$

If $\mathrm{CD}_{\alpha(n),\beta(n)}(f(n), g(n))$ then $\mathrm{SD}_{\alpha(n)+\beta(n)}(f(n), g(n))$.

$2 \to 1$

If $\mathrm{SD}_{\lambda(n)}(f(n), g(n)\log g(n))$ then $\mathrm{CD}_{\omega(\lambda(n)),\lambda(n)}(f(n), g(n))$.

Equivalent reformulations

$1 \to 2$

If $\mathrm{CD}_{\alpha(n),\beta(n)}(f(n), g(n))$ then $\mathrm{SD}_{\alpha(n)+\beta(n)}(f(n), g(n))$.

$2 \to 1$

If $\mathrm{SD}_{\lambda(n)}(f(n), g(n) \log g(n))$ then $\mathrm{CD}_{\omega(\lambda(n)),\lambda(n)}(f(n), g(n))$.

Samplable distributions hierarchy

WATSON, 2013

For any $a > 0$, $k > 0$ and $\epsilon > 0$ there is $D \in$ **PSamp** such that for every $F \in$ **Samp**(n^a), for infinitely many n the statistical distance between D_n and F_n is at least $1 - \frac{1}{k} - \epsilon$.

In previous notation: $\mathrm{SD}_{\frac{1}{k}+\epsilon}(\mathrm{poly}(n), n^k)$.

ITSYKSON, KNOP, SOKOLOV, 2015

For every a, b, c such that $0 < a < b$ and $c > 0$ there is $D \in$ **Samp**$(n^{\log^b n})$ such that for every $F \in$ **Samp**$(n^{\log^a n})$, for infinitely many n the statistical distance between D_n and F_n is at least $1 - \frac{1}{2^{(\log\log\log n)^c}}$.

In previous notation: $\mathrm{SD}_{\frac{1}{2^{(\log\log\log n)^c}}}(n^{\log^b n}, n^{\log^a n})$.

Samplable distributions hierarchy

WATSON, 2013

For any $a > 0$, $k > 0$ and $\epsilon > 0$ there is $D \in$ **PSamp** such that for every $F \in$ **Samp**(n^a), for infinitely many n the statistical distance between D_n and F_n is at least $1 - \frac{1}{k} - \epsilon$.
In previous notation: $\mathrm{SD}_{\frac{1}{k}+\epsilon}(\mathrm{poly}(n), n^k)$.

ITSYKSON, KNOP, SOKOLOV, 2015

For every a, b, c such that $0 < a < b$ and $c > 0$ there is $D \in$ **Samp**$(n^{\log^b n})$ such that for every $F \in$ **Samp**$(n^{\log^a n})$, for infinitely many n the statistical distance between D_n and F_n is at least $1 - \frac{1}{2^{(\log \log \log n)^c}}$.

In previous notation: $\mathrm{SD}_{\frac{1}{2^{(\log \log \log n)^c}}}(n^{\log^b n}, n^{\log^a n})$.

Samplable distributions hierarchy

WATSON, 2013

For any $a > 0$, $k > 0$ and $\epsilon > 0$ there is $D \in$ **PSamp** such that for every $F \in \textbf{Samp}(n^a)$, for infinitely many n the statistical distance between D_n and F_n is at least $1-\frac{1}{k} - \epsilon$.
In previous notation: $\mathrm{SD}_{\frac{1}{k}+\epsilon}(\mathrm{poly}(n), n^k)$.

ITSYKSON, KNOP, SOKOLOV, 2015

For every a, b, c such that $0 < a < b$ and $c > 0$ there is $D \in \textbf{Samp}(n^{\log^b n})$ such that for every $F \in \textbf{Samp}(n^{\log^a n})$, for infinitely many n the statistical distance between D_n and F_n is at least $1-\frac{1}{2^{(\log\log\log n)^c}}$.

In previous notation: $\mathrm{SD}_{\frac{1}{2^{(\log\log\log n)^c}}}(n^{\log^b n}, n^{\log^a n})$.

Samplable distributions hierarchy

WATSON, 2013

For any $a > 0$, $k > 0$ and $\epsilon > 0$ there is $D \in \textbf{PSamp}$ such that for every $F \in \textbf{Samp}(n^a)$, for infinitely many n the statistical distance between D_n and F_n is at least $1 - \frac{1}{k} - \epsilon$.
In previous notation: $\mathrm{SD}_{\frac{1}{k}+\epsilon}(\mathrm{poly}(n), n^k)$.

ITSYKSON, KNOP, SOKOLOV, 2015

For every a, b, c such that $0 < a < b$ and $c > 0$ there is $D \in \textbf{Samp}(n^{\log^b n})$ such that for every $F \in \textbf{Samp}(n^{\log^a n})$, for infinitely many n the statistical distance between D_n and F_n is at least $1 - \frac{1}{2^{(\log\log\log n)^c}}$.
In previous notation: $\mathrm{SD}_{\frac{1}{2^{(\log\log\log n)^c}}}(n^{\log^b n}, n^{\log^a n})$.

Proof of the Watson theorem for k = 2

1. Let A_1, ..., A_n, ...is an enumeration of all algorithms such that each algorithm occurred infinitely many times.
2. Consider sequences n_i, n_i^* such that $n_1 = 1$, $n_{i+1} = n_i^* + 1$ and $n_i^* = 2^{n_i^{a+1}}$
3. Consider the following algorithm (on input 1^n):
   - find i such that $n_i \leq n \leq n_i^*$;
   - if $n = n_i^*$ return $b \in \{0, 1\}$ such that $\Pr[A_i(1^{n_i}) = b] \leq \frac{1}{2}$;
   - else run $A_i(1^{n+1})$ $\frac{8 \log \epsilon}{\epsilon^2}$ times and return majority of answers.

Proof of the samplable distributions hierarchy

LIST DECODING

There is polynomial-time algorithm $C^{\bullet}(n, i, \lambda, \delta)$ such that if $\mathrm{supp}(\gamma) = \{0, \ldots, 2^n - 1\}$ and there is t such that $\Pr[\gamma = t] \geq \lambda$ then there is $i \leq (1 + \frac{1}{\lambda})^2$ such that $\Pr[C^{\gamma}(n, i, \delta, \lambda) = t] \geq 1 - \delta$.

Let us consider the following algorithm $C^{\gamma}(n, i, \lambda, \delta)$:

1. Let $k = \lceil \frac{1}{\lambda} + 1 \rceil$ and $\epsilon = \frac{\lambda^3}{10k}$;
2. We interpret i as a pair (a, b), where $a, b \in [k]$;
3. Request the oracle for $N = \lceil \frac{2(n+1+\log\frac{1}{\delta})}{\epsilon^2} \rceil$ samples of γ;
4. Consider the list y_1, ..., y_s of all elements with frequency at least $\lambda - \epsilon a$;
5. Return y_b if $b \leq s$ or 0 otherwise.

Proof of the samplable distributions hierarchy

LIST DECODING

There is polynomial-time algorithm $C^\bullet(n, i, \lambda, \delta)$ such that if $\mathsf{supp}(\gamma) = \{0, \dots, 2^n - 1\}$ and there is t such that $\Pr[\gamma = t] \geq \lambda$ then there is $i \leq (1 + \frac{1}{\lambda})^2$ such that $\Pr[C^\gamma(n, i, \delta, \lambda) = t] \geq 1 - \delta$.

Let us consider the following algorithm $C^\gamma(n, i, \lambda, \delta)$:

1. Let $k = \lceil \frac{1}{\lambda} + 1 \rceil$ and $\epsilon = \frac{\lambda^3}{10k}$;
2. We interpret i as a pair (a, b), where $a, b \in [k]$;
3. Request the oracle for $N = \lceil \frac{2(n+1+\log\frac{1}{\delta})}{\epsilon^2} \rceil$ samples of γ;
4. Consider the list y_1, ..., y_s of all elements with frequency at least $\lambda - \epsilon a$;
5. Return y_b if $b \leq s$ or 0 otherwise.

Proof of the samplable distributions hierarchy

LIST DECODING

There is polynomial-time algorithm $C^{\bullet}(n, i, \lambda, \delta)$ such that if $\mathrm{supp}(\gamma) = \{0, \ldots, 2^n - 1\}$ and there is t such that $\Pr[\gamma = t] \geq \lambda$ then there is $i \leq (1 + \frac{1}{\lambda})^2$ such that $\Pr[C^{\gamma}(n, i, \delta, \lambda) = t] \geq 1 - \delta$.

Let us consider the following algorithm $C^{\gamma}(n, i, \lambda, \delta)$:

1. Let $k = \lceil \frac{1}{\lambda} + 1 \rceil$ and $\epsilon = \frac{\lambda^3}{10k}$;
2. We interpret i as a pair (a, b), where $a, b \in [k]$;
3. Request the oracle for $N = \lceil \frac{2(n+1+\log \frac{1}{\delta})}{\epsilon^2} \rceil$ samples of γ;
4. Consider the list y_1, ..., y_s of all elements with frequency at least $\lambda - \epsilon a$;
5. Return y_b if $b \leq s$ or 0 otherwise.

Proof of the samplable distributions hierarchy

LIST DECODING

There is polynomial-time algorithm $C^{\bullet}(n, i, \lambda, \delta)$ such that if $\mathsf{supp}(\gamma) = \{0, \ldots, 2^n - 1\}$ and there is t such that $\Pr[\gamma = t] \geq \lambda$ then there is $i \leq (1 + \frac{1}{\lambda})^2$ such that $\Pr[C^{\gamma}(n, i, \delta, \lambda) = t] \geq 1 - \delta$.

Let us consider the following algorithm $C^{\gamma}(n, i, \lambda, \delta)$:

1. Let $k = \lceil \frac{1}{\lambda} + 1 \rceil$ and $\epsilon = \frac{\lambda^3}{10k}$;
2. We interpret i as a pair (a, b), where $a, b \in [k]$;
3. Request the oracle for $N = \lceil \frac{2(n+1+\log \frac{1}{\delta})}{\epsilon^2} \rceil$ samples of γ;
4. Consider the list y_1, ..., y_s of all elements with frequency at least $\lambda - \epsilon a$;
5. Return y_b if $b \leq s$ or 0 otherwise.

Proof of the samplable distributions hierarchy

LIST DECODING

There is polynomial-time algorithm $C^{\bullet}(n, i, \lambda, \delta)$ such that if $\mathrm{supp}(\gamma) = \{0, \dots, 2^n - 1\}$ and there is t such that $\Pr[\gamma = t] \geq \lambda$ then there is $i \leq (1 + \frac{1}{\lambda})^2$ such that $\Pr[C^{\gamma}(n, i, \delta, \lambda) = t] \geq 1 - \delta$.

Let us consider the following algorithm $C^{\gamma}(n, i, \lambda, \delta)$:

1. Let $k = \lceil \frac{1}{\lambda} + 1 \rceil$ and $\epsilon = \frac{\lambda^3}{10k}$;
2. We interpret i as a pair (a, b), where $a, b \in [k]$;
3. Request the oracle for $N = \lceil \frac{2(n+1+\log \frac{1}{\delta})}{\epsilon^2} \rceil$ samples of γ;
4. Consider the list y_1, ..., y_s of all elements with frequency at least $\lambda - \epsilon a$;
5. Return y_b if $b \leq s$ or 0 otherwise.

Magic tree

There exists a family of trees T_i such that

1. The set of vertices of T_i is a subset of $\{n_i, n_i + 1, \ldots, n_i^*\}$.
2. n_i^* is the root of T_i.
3. All leaves of T_i have numbers at most $m_i = 2n_i$.
4. The depth of T_i is $d_i = 2\lceil \log \log n_i \rceil$.
5. If p is a parent of n then $p \leq n^{\log n}$.
6. There is an algorithm that for any vertex n of T_i outputs the parent p of n and the number of children of p that are less than n in $poly(n)$ steps.
7. For every inner vertex v of T_i, v has $k = \lceil \frac{1}{\lambda(n_i^*)} + 1 \rceil^2$ children.