All Sampling Methods Produce Outliers

Samuel Epstein*

February 21, 2024

Abstract

This paper contains a simple proof of the sampling theorem in [Eps21] with exponentially improved bounds. A sampling method A is a probabilistic function that maps an integer N with probability 1 to a set containing N different strings. In the limit, greater outliers are guaranteed to exist in the output of A.

1 Discrete Sampling Theorem

A sampling method A is a probabilistic function that maps an integer N with probability 1 to a set containing N different strings. Let $P = P_1, P_2, \ldots$ be a sequence of measures over strings. For example, one may choose $P_1 = P_2 \ldots$ or choose P_n to be the uniform measure over n-bit strings. A conditional probability bounded P-test is a function $t : \{0,1\}^* \times \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that for all $n \in \mathbb{N}$ and positive real number r, we have $P_n(\{x : t(x|n) \geq r\}) \leq 1/r$. If $P_1, P_2, \ldots$ is uniformly computable, then there exists a lower-semicomputable such P-test t that is "maximal" (i.e., for which $t' \leq O(t)$ for every other such test t'). We fix such a t, and let $\overline{\mathbf{d}}_n(x|P) = \log t(x|n)$.

Lemma 1 *Let P be a computable measure on strings and let A be a sampling method. For all integers M and N, there exists a finite set $S \subset \{0,1\}^*$ such that $P(S) \leq 2M/N$, and with probability strictly more than $1 - 2e^{-M}$: $A(N)$ intersects S.*

Proof. We show that some possibly infinite set S satisfies the conditions, and thus, some finite subset also satisfies the conditions due to the strict inequality. We use the probabilistic method: we select each string to be in S with probability M/N and show that 2 conditions are satisfied with positive probability. The expected value of $P(S)$ is M/N. By the Markov inequality, the probability that $P(S) > 2M/N$ is at most $1/2$. For any set D containing N strings, the probability that S is disjoint from D is

$$(1 - M/N)^N < e^{-M}.$$

Let Q be the measure over N-element sets of strings generated by the sampling algorithm $A(N)$. The left-hand side above is equal to the expected value of

$$Q(\{D : D \text{ is disjoint from } S\}).$$

Again by the Markov inequality, with probability greater than $1/2$, this measure is less than $2e^{-M}$. By the union bound, the probability that at least one of the conditions is violated is less than $1/2 + 1/2$. Thus, with positive probability a required set is generated, and thus such a set exists. $\square$

*JP Theory Group. samepst@jptheorygroup.org

Theorem 1 *Let $P = P_1, P_2 \dots$ be a uniformly computable sequence of measures on strings and let A be a sampling method. There exists $c \in \mathbb{N}$ such that for all n and k:*

$$\Pr\left(\max_{a \in A(2^n)} \overline{\mathbf{d}}_n(a|P) > n - k - c\right) \geq 1 - 2e^{-2^k}.$$

Proof. We now fix a search procedure that on input N and M finds a set $S_{N,M}$ that satisfies the conditions of Lemma 1. Let $t'(a|n)$ be the maximal value of $2^n/2^{k+2}$ such that $a \in S_{2^n,2^k}$ for some integer k. By construction, t' is a computable probability bound test, because $P(\{x : t'(x|n) = 2^\ell\}) \leq 2^{-\ell-1}$, and thus $P(t'(x|n) \geq 2^\ell) \leq 2^{-\ell-1} + 2^{-\ell-2} + \dots$ With the given probability, the set $A(2^n)$ intersects $S_{2^n,2^k}$. For any number a in the intersection, we have $t'(x|n) \geq 2^{n-k-2}$, thus by the optimality of t and definition of $\overline{\mathbf{d}}$, we have $\overline{\mathbf{d}}_n(a|P) > n - k - O(1)$. □

An incomplete sampling method A takes in a natural number N and outputs, with probability $f(N)$, a set of N numbers. Otherwise A outputs $\perp$. f is computable.

Corollary 1 *Let $P = P_1, P_2 \dots$ be a uniformly computable sequence of measures on strings and let A be an incomplete sampling method. There exists $c \in \mathbb{N}$ such that for all n and k:*

$$\Pr_{D=A(n)}\left(D \neq \perp \text{ and } \max_{a \in D} \overline{\mathbf{d}}_n(a|P) \leq n - k - c\right) < 2e^{-2^k}.$$

2 Continuous Sampling Method

Let $\mu = \mu_1, \mu_2, \dots$ be a uniformly computable sequence of measures over infinite sequences. Similar way as for strings in the introduction, the randomness deficiency $\overline{\mathbf{D}}_n(\omega|\mu)$ for sequences ω is defined using lower-semicomputable functions $\{0,1\}^\infty \times \mathbb{N} \to \mathbb{R}_{\geq 0}$. A continuous sampling method C is a probabilistic function that maps, with probability 1, an integer N to an infinite encoding of N different sequences.

Theorem 2 *There exists $c \in \mathbb{N}$ where for all n:*

$$\Pr\left(\max_{\alpha \in C(2^n)} \overline{\mathbf{D}}_n(\alpha|\mu) > n - k - c\right) \geq 1 - 2.5e^{-2^k}.$$

Proof. For $D \subseteq \{0,1\}^\infty$, $D_m = \{\omega[0..m] : \omega \in D\}$. Let $g(n) = \arg\min_m \Pr_{D=C(n)}(|D_m| < n) < 0.5e^{-2^n}$ be the smallest number m such that the initial m-segment of $C(n)$ are sets of n strings with very high probability. g is computable, because C outputs a set of distinct infinite sequences with probability 1. For probability ψ over $\{0,1\}^\infty$, let $\psi^m(x) = [|x| = m]\psi(\{\omega : x \sqsubset \omega\})$. Let $\mu^g = \mu_1^{g(1)}, \mu_2^{g(2)}, \dots$ be a uniformly computable sequence of discrete probability measures and let A be a discrete incomplete sampling method, where for random seed $\omega \in \{0,1\}^\infty$, $A(n,\omega) = C(n,\omega)_{g(n)}$

if $|C(n,\omega)_{g(n)}| = n$; otherwise $A(n,\omega) = \perp$. So $\Pr[A(n) = \perp] < 0.5e^{-2^n}$.

$$
\begin{aligned}
&\Pr\left(\max_{\alpha \in C(2^n)} \overline{\mathbf{D}}_n(\alpha|\mu) \leq n - k - O(1)\right) \\
\leq &\Pr_{Z=C(2^n)}\left((|Z_{g(n)}| < 2^n) \text{ or } (|Z_{g(n)}| = 2^n \text{ and } \max_{\alpha \in Z} \overline{\mathbf{D}}_n(\alpha|\mu) \leq n - k - O(1)\right) \\
\leq &\Pr_{D=A(2^n)}\left(D = \perp \;\text{ or }\; (D \neq \perp \;\text{ and }\; \max_{x \in D} \overline{\mathbf{d}}_n(x|\mu^g) \leq n - k - O(1))\right) \\
< &0.5e^{-2^n} + 2e^{-2^k} \qquad (1) \\
\leq &2.5e^{-2^k},
\end{aligned}
$$

where Equation 1 is due to Corollary 1. □

3 Output of Randomized Algorithms

In this section, we prove that the non-automatic output of randomized algorithms are guaranteed to have high $\mathbf{D}$ scores, i.e. be outliers. Let $\lambda = \lambda_1, \lambda_2, \ldots$ and $\mu = \mu_1, \mu_2, \ldots$ be uniformly computable sequences of measures over infinite sequences. Each λ_n is non-atomic.

Theorem 3 *There is a constant $f \in \mathbb{N}$, dependent on μ and λ, where for all $n \in \mathbb{N}$, $\lambda_n\left\{\alpha : \overline{\mathbf{D}}_n(\alpha|\mu) > n - f\right\} > 2^{-n-f}$.*

Proof. We define the continuous sampling method C, where on input n, randomly samples n elements from λ_n. Let $d_n = \lambda_n\{\alpha : \overline{\mathbf{D}}_n(\alpha|\mu) > n - b\}$, where b is the constant in Theorem 2. Evoking this theorem, with $k = 0$,

$$
\begin{aligned}
\Pr\left(\max_{\alpha \in C(2^n)} \overline{\mathbf{D}}_n(\alpha|\mu) > n - b\right) &> 1 - 2.5e^{-1} \\
1 - (1 - d_n)^{2^n} &> 1 - 2.5e^{-1} \\
1 - 2^n d_n &< 2.5/e \\
d_n &> (1 - 2.5/e)2^{-n} \\
\lambda_n\{\alpha : \overline{\mathbf{D}}_n(\alpha|\mu) > n - b\} &> 2^{-n-c} \\
\lambda_n\{\alpha : \overline{\mathbf{D}}_n(\alpha|\mu) > n - f\} &> 2^{-n-f}.
\end{aligned}
$$

□

4 Necessity of Double Exponential

Theorem 1 showed that the probability that $A(2^n)$ contains no strings of randomness deficiency less than $n - k$ decreases double exponentially in k. We show that at least a double exponential probability is required for $k = n - O(1)$. Let P_n be the uniform measure on $(n+2)$-bit strings. The algorithm A that on input 2^n generates a random set of 2^n strings of length $n + 2$ satisfies

$$\Pr\left(\forall x \in A(2^n) : \overline{\mathbf{d}}_n(x|P) \leq 2\right) \geq 2^{-2^n}.$$

The reasoning for this is as follows. For at most a quarter of the $(n+2)$-bit strings, we have $\overline{\mathbf{d}}_n(x|P) \geq 3$, by definition of a probability bounded test t. A random selection of $N = 2^n$ different

$(n+2)$-bit strings, contains no such string with a probability of at least 2^{-N}. We consider the following situation. In a bag with $4N$ balls, N balls are marked. One selects N balls one by one. We consider the probability that no marked ball is drawn if previously no marked ball was drawn. The smallest probability appears at the last draw when there are $T = 4N - (N-1)$ balls in the bag. This probability is $(T-N)/T \geq 1/2$.

5 Partial Sampling Methods

A partial sampling method is a sampling method that can output with probability less than 1. Theorem 1 does not hold for partial sampling methods B. Let P_n be the uniform measure on $(n+1)$-bit strings. Let $\#B(N)$ represent the event that B halts and outputs a set of size N. We present a partial sampling method B for which

$$\Pr\left(\#B(2^n) \text{ and } \forall x \in B(2^n) : \overline{\mathbf{d}}_n(x|P) \leq 1\right) \geq 2^{-n}.$$

For at most half of the $(n+1)$-bit strings, we have $\overline{\mathbf{d}}_n(x|P) \geq 2$. On input 2^n,the partial sampling method B generates a random natural number s bounded by 2^n, searches for s strings x of length $n+1$ with $\overline{\mathbf{d}}_n(x|P) \geq 2$, and outputs 2^n other $(n+1)$-bit strings. For some s, this search may never terminate. If A chooses to be precisely equal to the number of strings satisfying the condition, then it outputs only strings with deficiency at most 1, and the claim is proven. However partial sampling methods do exhibit the following properties

Theorem 4 *Let $P = P_1, P_2, \ldots$ be a uniformly computable sequence of measures and B be a partial sampling method, where $\#B(N)$ represents the event that $B(N)$ terminates and outputs a set of N strings.*

$$\Pr\left(\#B(N) \text{ and } \forall x \in B(2^n) : \overline{\mathbf{d}}_n(x|P) \leq n-k\right) \leq O(k2^{-k}).$$

Proof. Let Q be the lower-semicomputable semimeasure over sets of size 2^n such that $Q(D)$ equals the probability that $B(N) = D$. We show that

$$\Pr\left(\#B(N) \text{ and } \forall x \in B(2^n) : \overline{\mathbf{d}}_n(x|P) \leq n-k+\log k + O(1)\right) \leq O(2^{-k}).$$

This result is followed by a redefinition of k. We write Q as a uniform mixture over at most 2^k measures Q_i with finite support, and one lower semi-computable semimeasure Q_*:

$$Q = 2^{-k}\left(Q_1 + Q_2 + \ldots Q_f + Q_*\right).$$

With $f \leq 2^k$, we assume that the finite descriptions of $Q_1, \ldots, Q_f$ are enumerated one by one by a program (that may never terminate). For each enumerated measure Q, we search for a set S_i that satisfies the conditions of Lemma 1 for $M = k$. Let $S = \bigcup_{i \leq f} S_i$. Also, $P(S) \leq k2^{k+1-n}$; thus every element in S satisfies $\overline{\mathbf{d}}_n(x|P) \geq n-k+\log k + O(1)$.

The probability that $A(2^n)$ produces a set that does not contain such an element is at most $2^{-k} + 2e^{-k}$ because we can equivalently generate a set D by randomly selecting j from the list $[1, \ldots, f, *, \infty]$ with probabilities $[2^{-k}, \ldots, 2^{-k}, 2^{-k}r, 1-(f+r)2^{-k}]$ and generating a random set D from Q_j if $j \neq \infty$ and letting D be undefined otherwise. The probability that D is defined and does not contain an element from S is at most the probability $j = *$, which is $\leq 2^{-k}$, plus the probability that $j \in \{1, \ldots, f\}$ times $2e^{-k}$. □

References

[Eps21] Samuel Epstein. All sampling methods produce outliers. *IEEE Transactions on Information Theory*, 67(11):7568–7578, 2021.