LIMITATIONS ON AUTOREGRESSIVE TOKEN GENERATION VIA SEQUENTIAL SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study autoregressive token generation through the notion of sequential samplers: A circuit that samples the next symbol only given the previous ones and possibly a small amount of state. We show that (1) a sequential sampler with k bits of state can be simulated by a stateless one at a roughly $O(2^k)$ size overhead; (2) efficient insertional sampling (Stern, Chan, Kiros, Uszkoreit, ICML 2019) is more powerful than efficient sequential sampling in any fixed order; (3) sequential samplers with small state cannot establish certain forms of covert cryptographic channels. Result (1), together with the pseudoentropy vs. sequential sampleability duality of Vadhan and Zheng (STOC 2012), characterizes efficient sequential sampleability with small state. Result (3) provides a potential safety mechanism against subliminal coordination among malicious LLM agents.

1 INTRODUCTION

Autoregressive language models generate an entire response through a sequence of local decisions: sample the next token given the prompt and the tokens already produced. An ordinary randomized sampler, in contrast, is unconstrained by the order in which its output is produced. A long pseudorandom string, for example, can be derived from a short random seed, even though the visible prefix does not reveal its continuation.

In this work we fully characterize the class of distributions that can be sampled *efficiently* one token at a time, using only the preceding output.

The probability chain rule factors *every* distribution into next-token conditionals: $p(x_1x_2x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1x_2)$. However, this formula says nothing about the complexity of sampling those conditionals. An efficient algorithm for generating a whole string need not provide an efficient way to continue from a prefix drawn from the same distribution. Impagliazzo & Levin (1990) furnish a sufficient hypothesis for *universal* efficient sequential sampleability: The average-case invertibility of all efficient functions. Their hypothesis is unreasonable as it would forbid cryptography. (We may be in for surprises!)

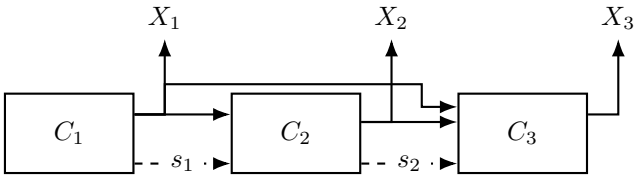


Figure 1: A three-token sequential sampler.

We study an abstraction of autoregressive generation called a *sequential sampler* (see Fig. 1). This is a collection of probabilistic circuits $C(1^n) = (C_1, C_2, \ldots, C_n)$ generating token sequence $X_1, X_2, \ldots, X_n$. Circuit C_i takes as inputs the prefix $X_{<i} = (X_1, \ldots, X_{i-1})$, and possibly a bounded state s_{i-1} produced by C_{i-1}. It outputs the next token-state pair (X_i, s_i). When s_i are empty the sampler is *stateless*.

Every step can reread the entire prefix sampled so far. Private information generated during one step, including randomness, can only be explicitly passed along through the state. Prompts, as well as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

other fixed model parameters, can be hardwired into the circuits, so we impose no restriction on the training procedure, attention mechanism, or depth of the next-token computation beyond efficiency, i.e., that the circuit size be polynomial in the total output length. The model captures generation from fixed autoregressive parameters when all information retained between steps is determined by the prompt and generated prefix.

Earlier work by Lin et al. (2021) identifies sampleability limitations for autoregressive models satisfying the stronger assumption that next-token probabilities are efficiently computable. Fortnow (2026) shows that the distributions that can be sampled in this manner are precisely the ones whose *cumulative* distribution function is efficiently computable (Levin, 1986). Our model allows for a richer class of distributions.

More relevant to our model is the limitation observed by Wu (2023): A stateless sequential sampler cannot generate a number followed by its prime factorization, unless factoring is easy. Vadhan & Zheng (2012) and Agrawal et al. (2019) fully characterize this type of obstruction for stateless, fixed-order samplers.

Our approach is based on the duality between conditional sampleability and *pseudoentropy* discovered by Vadhan & Zheng (2012). Agrawal et al. (2019) interpret this connection in the context of stateless sequential sampling. We derive analogous characterizations for stateful and *insertional* samplers.

1.1 Our Results

All reasonable sampling notions tolerate some error of the sampler $C(1^n)$ relative to the target distribution Y. A natural measure, which we adopt in this work, is their KL-divergence $D_{\mathsf{KL}}\left(Y \parallel C(1^n)\right)$. Note that small KL-divergence implies small total variation distance.

Characterizing sequential samplers. For left-to-right generation, the relevant quantity is *next-block pseudoentropy*. It measures how much entropy the next token can appear to have, given the true preceding prefix, summed over the positions. The *next-block pseudoentropy gap* is the difference between this quantity and the true entropy of the complete string. This equivalence follows from the work of Vadhan & Zheng (2012) and was informally discussed by Agrawal et al. (2019). We give a formal, quantitative statement and proof:

Theorem 1.1 (Vadhan & Zheng (2012); Agrawal et al. (2019), see Theorem B.1)**.** *A distribution has a small next-block pseudoentropy gap if and only if it admits an efficient stateless sequential sampler with small KL error.*

The next-token circuits can be chosen with full knowledge of the target, independently of any learning algorithm. A large pseudoentropy gap therefore identifies a limitation that persists even when training and model selection are unrestricted.

We show that the above extends beyond stateless samplers. The reason is that private state can be eliminated, at a cost polynomial in the number of possible states:

Theorem 1.2 (Theorem 4.2, informal)**.** *A sequential sampler with k bits of state can be approximated in KL-divergence by a stateless sequential sampler. The circuit-size overhead is 2^k and polynomial in the other parameters, including token alphabet size.*

A factor 2^k blowup in overhead is inevitable for simulation that estimates the state distribution from independent samples, which seems necessary in general. As intuition, a pseudorandom generator with a k-bit seed admits an efficient stateful implementation that stores the seed. A stateless implementation, however, seems to effectively require brute-force search over all possible seeds. See Section C.3.

Insertional Samplers. An *insertional sampler* can insert a new token anywhere in the output, not just at the terminal position. The Insertion Transformer (Stern et al., 2019) is endowed with this capability. The sampler still retains bounded private state, but can generate information in an order suited to the task.

We introduce the notion of *insertional pseudoentropy gap* to measure how much the apparent conditional entropy can exceed the actual conditional entropy, summed over the insertions, for every

insertional decomposition of the target. We show a duality between insertional pseudoentropy and sampleability:

Theorem 1.3 (Theorem 5.4, informal)**.** *A distribution has a small insertional pseudoentropy gap if and only if it admits an efficient stateless insertional sampler with small KL error.*

Insertional samplers are more powerful than sequential ones: Going back to the factoring example, one can simply output the factors first, then insert their product in the first position of the sequence. In fact, this is not just a property of factoring: For any one-way function f,[1] the pair $(f(X), X)$ is hard to sample in a left-to-right manner but easy to sample insertionally (Proposition F.1).

In contrast to the pair $(f(X), X)$ which is easy to sample from right-to-left, a pseudorandom generator is not computationally samplable for *any fixed* ordering.[2] This is because it has next-bit pseudoentropy with respect to every fixed ordering of the output bits. Thus, one might expect that the pseudorandom generator output distribution is not insertionally sampleable. Yet, surprisingly, in Proposition 5.5 we demonstrate the existence of pseudorandom generators whose output can be sampled insertionally. Our construction crucially exploits the stochasticity of the insertion positions.

Nevertheless, we demonstrate a separation between insertional and arbitrary samplers.

Theorem 1.4 (Theorem F.2, informal)**.** *Assuming the existence of very hard-to-invert one-way functions, there is an efficiently sampleable distribution that cannot be sampled efficiently in a stateless insertional manner.*

The token alphabet in our hard distribution has bit-length linear in the sequence length n. We do not know if this dependence is necessary.

Consequences (Section 6). Our results reveal fundamental tradeoffs between the expressive power of generative models and their cryptographic capabilities. More restrictive architectures provide automatic safety guarantees. More expressive architectures, in contrast, could be useful for training cryptographic functionalities.

Cryptographic capabilities and safeguards. Autoregressive models face limits even on useful tasks that ordinary programs perform routinely. Our results imply that they cannot faithfully sample the output of a pseudorandom generator, a necessity for encryption. Supporting these capabilities requires sufficient private state or access to external resources.

This limitation can be beneficial for safety. Untrusted agents might try to establish a covert channel via steganographic techniques. Bounding the state of transformer-type agents constrains their ability to communicate in secret.

A case for insertion-based models. At the level of generation rules, insertion is strictly more expressive than left-to-right generation in our stateless setting. The freedom to choose where to write can therefore expand a model's capabilities without adding private memory. For example, as we have shown, useful cryptography can be computed statelessly by insertion-based models but not by models with any apriori fixed order. This provides a theoretical foundation for pursuing insertion-based models as an alternative to strictly autoregressive generation.

1.2 Further Related Work

Computational entropy provides several ways to express the difficulty of producing correlated outputs. Inaccessible entropy, introduced by Haitner et al. (2009), measures the entropy an online generator can produce when its previous random coins are included in the conditioning information. Agrawal et al. (2019) unify this notion with next-block pseudoentropy through hardness in relative entropy. We use the prefix-only sampling connection from this theory as a foundation; the insertional analysis must additionally handle intermediate strings and the information in their generation order.

[1] Recall that a one-way function f is a function where $y = f(x)$ is easy to compute for every x, but it is hard to find a preimage of y when x is sampled uniformly at random.

[2] A pseudorandom generator $G\colon \{0,1\}^n \to \{0,1\}^m$ is a length-extending function where $G(s)$ for a uniformly random $s \leftarrow \{0,1\}^n$ is computationally indistinguishable from a uniformly sampled string over $\{0,1\}^m$.

Our next-token circuits sample their outputs, and their output probabilities need not be efficiently evaluable. This distinguishes the model from the probability-evaluation setting of Lin et al. (2021). Wu's factoring-based example (Wu, 2023) already exhibits a limitation for randomized prefix-only generation. The entropy characterization provides a general criterion; our insertional analysis extends this viewpoint to alternative generation orders.

Recent work also studies the computational role of generation order. Kim et al. (2025) analyze hard masking conditionals under computational assumptions and the benefits of choosing favorable orders. Yang et al. (2026) compare autoregressive, masked, and editing-based generation; their autoregressive simulation of arbitrary orders emits position–token pairs and permits postprocessing. Our samplers must directly produce the prescribed output string, so that simulation does not imply equivalence in our model. Jiang et al. (2026) study parallel sampling efficiency and simulation of randomized circuits by diffusion models with auxiliary tokens. Their resource bounds complement our focus on the distributions that can be generated under restrictions on private state and output order.

1.3 Paper Organization

Section 2 includes preliminaries needed for this paper. Section 3 contains the definition of sequential samplers. Section 4 describes our stateless simulation of stateful samplers. Section 5 defines insertional samplers and pseudoentropy and proves their equivalence. Section 6 discusses consequences of our results.

Section A contains further preliminaries. Section B proves a characterization of sequential samplers via next-block pseudoentropy. Section C and Section D contain full proofs and expanded statements for Section 4 and Section 5 respectively. Section E describes our PRG sampleable by insertional samplers. Section F describes separations between notions of sampling and pseudoentropy. Finally, Section G contains further proofs and discussion on consequences of our results, complementing Section 6. All appendices restate the theorems being proven.

2 Preliminaries

All logarithms are in base two. Write $[n] = \{1, \ldots, n\}$, U_m for the uniform distribution on $\{0,1\}^m$, and $x_{<i} = (x_1, \ldots, x_{i-1})$. A string is a subsequence of another if it can be obtained by deletions. Every randomized circuit uses fresh independent fair coins; circuit families may be nonuniform. A negligible function is eventually smaller than every inverse polynomial; a noticeable function is eventually larger than some inverse polynomial.

For finite distributions P, Q, write $\Delta(P, Q)$ for their total variation distance and $D_{\mathsf{KL}}(P \parallel Q) = \sum_x P(x) \log(P(x)/Q(x))$ for their KL-divergence, with value $+\infty$ if $Q(x) = 0 < P(x)$ for some x. Our approximation guarantees use D_{KL} (target $\parallel$ sampler). Pinsker's inequality gives $\Delta(P, Q) \le \sqrt{(\ln 2) D_{\mathsf{KL}}(P \parallel Q)/2}$. We write $H(X) = \mathbb{E}[-\log \Pr[X = x]]$ for Shannon entropy, with the expectation over $x \leftarrow X$, and $H(X \mid W) = \mathbb{E}_{w \leftarrow W} H(X \mid_{W=w})$ for conditional entropy.

Two distributions are (T, ε) indistinguishable (ind.) if every Boolean circuit of size at most T distinguishes them with advantage at most ε. Computational indistinguishability means that every polynomial-size circuit family has negligible advantage. Pseudoentropy measures the entropy of distributions that are indistinguishable in this sense. In particular, the next-block pseudoentropy of $Y = (Y_1, \ldots, Y_n)$ at parameters (T, ε) is at least k if, for an independent uniform $I \in [n]$, there is a jointly distributed Z over the token alphabet such that $(Y_{<I}, Z)$ is (T, ε) indistinguishable from $(Y_{<I}, Y_I)$ and $H(Z \mid Y_{<I}) \ge k/n$. Its gap is measured from $H(Y)$. Given visible information W, include W in both compared distributions and in the entropy conditioning; the gap is then measured from $H(Y \mid W)$. For families indexed by n, an asymptotic lower bound $k(n)$ requires (n^c, n^{-c}) pseudoentropy at least $k(n) - n^{-c}$ for every $c > 0$ and all sufficiently large n. Detailed preliminaries can be found in Section A.

3 Stateless Sequential Sampling Algorithms

We define sequential samplers:

Definition 3.1 ((ℓ, B)-Sequential Sampler)**.** *We define a (stateless) sequential sampling algorithm of size B to be a sequence of randomized boolean circuits $C(1^n) = (C_1, C_2, \ldots, C_n)$ for an output length $n \in \mathbb{N}$ such that all circuits C_i are of size at most B.*

We use $C(1^n)$ to denote the distribution of $(x_1, \ldots, x_n)$ sampled iteratively as follows: for $i = 1$, sample $x_1 \leftarrow C_1(1^n)$ and for $i > 1$ sample $x_i \leftarrow C_i(1^n, x_1, \ldots, x_{i-1})$. We use ℓ to denote the length of the x_i's.

All invocations use fresh independent random coins. The input 1^n is fixed for each circuit family and is suppressed when convenient.

4 Eliminating Private State

Private state lets a sampler retain information that its visible output has not revealed. We show how to recover the next-token distribution from the prefix alone, at a cost essentially linear in the number of possible states. The simulation estimates the distribution of the hidden state conditioned on the visible prefix.

Definition 4.1 (Sequential sampler with k bits of state)**.** *A sequential sampler with k bits of state and size B for n blocks over alphabet $\mathcal{X}$ is a sequence $C = (C_1, \ldots, C_n)$ of randomized circuits, each of size at most B, with state space $\mathcal{S} = \{0,1\}^k$. The variables $(X_1, S_1), \ldots, (X_n, S_n)$ are generated using independent random coins for each circuit:*

$$(X_1, S_1) \leftarrow C_1(1^n),$$

and for $i \geq 2$,

$$(X_i, S_i) \leftarrow C_i(1^n, X_{<i}, S_{i-1}).$$

The sampler C outputs $X = (X_1, \ldots, X_n)$; the states $S_1, \ldots, S_n$ are not part of its output.

When $k = 0$, this is a stateless sequential sampler. The following theorem bounds the cost of removing state while approximating the output distribution.

Theorem 4.2 (State elimination for sequential samplers (small KL))**.** *Let $n \geq 1$ and $k, \ell \geq 0$ be integers, and let $0 < \delta \leq 1$. Let $C = (C_1, \ldots, C_n)$ be a sequential sampler with k bits of private state and size at most B per circuit, over alphabet $\mathcal{X} = \{0,1\}^\ell$, and put $\mathcal{S} = \{0,1\}^k$. There is a randomized stateless sequential sampler $D = (D_1, \ldots, D_n)$ such that*

$$D_{\mathsf{KL}}\left(C(1^n) \parallel D(1^n)\right) \leq \delta$$

and every D_i has Boolean circuit size at most

$$O\Big(\frac{n^3|\mathcal{X}||\mathcal{S}|}{\delta}\left[B + \log^3\Big(\frac{2n|\mathcal{X}||\mathcal{S}|}{\delta}\Big)\right]\Big).$$

The circuit construction is deterministic; all sampling and estimation randomness is internal to the circuits.

Proof overview. Put $a = |\mathcal{X}|$, $m = |\mathcal{S}|$, and $M = \lceil 4n^2 am/\delta \rceil$. For a list L of at most M states, let its smoothed distribution assign each $s \in \mathcal{S}$ probability $(1 + \#_L(s))/(|L| + m)$, where $\#_L(s)$ counts occurrences of s in L. Each draw below uses a fixed-precision implementation with at most $1/M$ additional KL error.

Let D_1 output the first component of $C_1(1^n)$. For $i \geq 2$, D_i receives $x_{<i}$ and performs the following steps using fresh independent coins:

1. Draw M pairs from $C_1(1^n)$. Let L_1 contain the states of those pairs whose output token equals x_1.
2. For $j = 2, \ldots, i-1$, draw M independent states from the smoothed distribution of L_{j-1} and apply $C_j(1^n, x_{<j}, \cdot)$ to each. Retain in L_j the states whose output token equals x_j.
3. Draw a state s from the smoothed distribution of L_{i-1} and output the token produced by $C_i(1^n, x_{<i}, s)$.

All lists are recomputed inside D_i, so no private information passes between transitions.

The analysis estimates the joint distribution of a token and state *before* filtering by the token. Applying the KL chain rule to its Laplace estimate bounds the average error after conditioning, even when some tokens are rare. If ε_j is the expected KL error of the state estimate after j tokens, this gives $\varepsilon_j \le \varepsilon_{j-1} + \gamma + \beta$, where $\varepsilon_0 = 0$, $\gamma = \log((M + am)/(M + 1))$, and $\beta \le 1/M$. Data processing, convexity, and the KL chain rule then give $D_{\mathsf{KL}}\left(C(1^n) \parallel D(1^n)\right) \le \sum_{j<n} \varepsilon_j \le \binom{n}{2}(\gamma + \beta) \le \delta$.

To obtain the stated circuit size, all M draws from one smoothed list are implemented together using sorting networks, in $O(M \log^3(2Mm))$ gates. Each of the at most n stages additionally uses M copies of a transition circuit, giving the claimed bound. The full construction and proof appear in Theorem C.4. □

For polynomial-size alphabets and $k = O(\log n)$, the simulation has polynomial size at any specified inverse-polynomial KL accuracy. Thus every distribution generated efficiently with logarithmic private state can be approximated efficiently without it. Together with the known characterization by next-block pseudoentropy (Theorem B.1), this extends sampling obstructions to logarithmic state; the cryptographic consequences below apply this connection.

5 Insertional Sampling

An insertional sampler chooses both the next token and where to place it. This abstracts the insertion framework of Stern et al. (2019), with one sampled insertion per step. The current string is the only information retained between steps; each circuit uses fresh independent coins.

Definition 5.1 (Stateless Insertional Sampler). *Let $n, \ell \ge 1$ and let L be a distribution over $(\{0,1\}^\ell)^n$. An insertional sampler $C = (C_1, \dots, C_n)$ of L is a sequence of randomized (possibly unbounded) circuits, where*

- *$C(1^n) \coloneqq C_n(C_{n-1}(\cdots(C_1(\emptyset)))) \equiv L$, and*
- *for every $i \in [n]$ and $x \in (\{0,1\}^\ell)^{i-1}$, $C_i(x)$ is a distribution over strings in $(\{0,1\}^\ell)^i$ for which x is a subsequence.*

We say that C has size B if for all $i \in [n]$, C_i is of size at most B.

We characterize insertional sampling using *insertional pseudoentropy*. The possible intermediate strings replace the fixed prefixes of sequential sampling.

Definition 5.2 (Insertional decomposition). *Let $n, \ell \ge 1$ and let L be a distribution over $(\{0,1\}^\ell)^n$. An insertional decomposition $(L_0, \dots, L_n)$ of L is a sequence of jointly distributed random variables, where*

- *L_0 is the fixed distribution over the empty string $\emptyset$,*
- *$L_n \equiv L$, and*
- *for every $i \in [n]$ and $x \in (\{0,1\}^\ell)^{i-1}$ with $\Pr[L_{i-1} = x] > 0$, $L_i \mid L_{i-1} = x$ is a distribution over strings in $(\{0,1\}^\ell)^i$ for which x is a subsequence.*

Definition 5.3 (Insertional Pseudoentropy). *Let $n, \ell \ge 1$ and let L be distributed over $(\{0,1\}^\ell)^n$. Let $T \ge 1$, $k \ge 0$, and $0 < \varepsilon < 1$; these quantities may depend on a security parameter λ. We say that L has (T, ε) insertional pseudoentropy at least k if, for every insertional decomposition $(L_0, \dots, L_n)$ of L, with I uniform on $[n]$ independently of the decomposition, there exists a random variable Z, jointly distributed with $(I, L_0, \dots, L_n)$ and satisfying $Z \in (\{0,1\}^\ell)^i$ whenever $I = i$, such that*

1. *$H(Z \mid I, L_{I-1}) - H(L_I \mid I, L_{I-1}) \ge \frac{k - H(L)}{n}$, and*
2. *(I, L_I, L_{I-1}) and (I, Z, L_{I-1}) are (T, ε) nonuniformly indistinguishable.*

The insertional pseudoentropy is the supremum of the values k satisfying this definition.

The insertional pseudoentropy gap is the insertional pseudoentropy minus $H(L)$. Subtracting the actual conditional entropy prevents randomness in the insertion order from being counted as a gap.

Quantifying over every decomposition accounts for the sampler's freedom to choose where to insert. The following theorem characterizes efficient insertional sampling; the dependence on alphabet size is polynomial.

Theorem 5.4 (Characterization of stateless insertional sampling)**.** *There exist a universal constant $c_0 \in (0, 1]$ and a polynomial p with nonnegative coefficients and constant term at least one such that the following holds. Let $n, \ell \geq 1$ and let L take values in $(\{0,1\}^\ell)^n$. Let $t \geq 1$, $0 < \epsilon, \eta, \sigma < 1$, and $\delta \geq \epsilon$.*

1. *If L has (t, ϵ) insertional pseudoentropy at least $H(L)+\delta$, then, for every $B \geq 1$ satisfying*

$$B \leq \frac{\min\{t^{c_0},\, \sigma/(n\epsilon)\}}{p(n, 2^\ell, 1/\eta, \log(n/\sigma))},$$

every size-B stateless insertional sampler C over $\{0,1\}^\ell$ satisfies

$$D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) > \delta - \sigma - n\eta.$$

2. *If the (t, ϵ) insertional pseudoentropy of L is at most $H(L) + \delta - \epsilon$, then there exists a size-B stateless insertional sampler C over $\{0,1\}^\ell$ with*

$$B \leq p(t, n, 2^\ell, 1/\epsilon), \qquad D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) \leq \delta.$$

The proof relates final KL-divergence to conditional sampling along an insertional decomposition, and then applies the conditional pseudoentropy characterization. The complete argument appears in Theorem D.6.

Finally, insertion does not recover all efficient sampling: Theorem F.2 proves a separation under $2^{\gamma\lambda^\alpha}$-secure one-way functions, for constants $\gamma > 0$ and $1/2 < \alpha < 1$. This construction uses $2t$ tokens over an alphabet of size $2t\, 2^{2t}$; the definitions and proofs are given in Sections D.1 and F.

5.1 An insertionally samplable PRG

Insertion can help simply by changing the generation order. For a one-way function f, the distribution $(f(X), X)$ for uniform X is hard to sample sequentially, but an insertional sampler can generate X first and insert $f(X)$ before it (Proposition F.1). We give a stronger example: one distribution that is easy to sample insertionally but hard in *every fixed output order*.

A *pseudorandom generator* (PRG) is a polynomial-time family $G : \{0,1\}^k \to \{0,1\}^m$, with $m > k$, whose output on a uniform seed is computationally indistinguishable from U_m. Here security is against nonuniform polynomial-size circuits: every such circuit has negligible distinguishing advantage. The following construction shows that an insertional sampler can generate pseudorandom bits without retaining the seed privately.

Proposition 5.5 (Insertionally samplable pseudorandom generators)**.** *Let $G : \{0,1\}^t \to \{0,1\}^{2t^2}$ be a polynomial-time pseudorandom generator secure against nonuniform polynomial-size circuits. There exists a polynomial-time pseudorandom generator $H : \{0,1\}^{3t^2} \to \{0,1\}^{4t^2}$ whose output distribution admits an exact polynomial-size stateless insertional bit sampler. The same holds for any apriori fixed ordering, rather than just left-to-right.*

Construction and proof idea. The sampler first generates bits from which it can recover a seed. After using this seed, it inserts additional random bits so that the final string conceals which bits were originally generated. The $3t^2$ random bits used in the following process form the seed of H.

1. Append t^2 independent uniform bits $X_1, \ldots, X_{t^2}$. Partition them into t consecutive groups of t bits, and let s_j be the parity of group j. Put $s = s_1 \cdots s_t$.
2. Append the $2t^2$ bits of $G(s)$, recomputing s from the first t^2 positions at each step.
3. For $i = 1, \ldots, t^2$, choose independent uniform bits b, r. Insert b immediately before X_i if $r = 0$, and immediately after it otherwise. Let Z_i be the resulting pair. The final output is $Z \| G(s)$, where $Z = Z_1 \| \cdots \| Z_{t^2}$.

This process is stateless: the string length determines the phase and step, and X_i occupies position $2i-1$ when its partner is inserted. All transitions have polynomial-size circuits.

The pairs Z_i are independent and uniform on $\{00, 01, 10, 11\}$. Given Z, the original bits X_i remain independent. An unequal pair leaves its original bit uniform, so a group containing an unequal pair has uniform parity even given Z. A group has only equal pairs with probability 2^{-t}. By a union bound, except with probability at most $t2^{-t}$, all t parities are independent and uniform given Z. Thus (Z, s) is within $t2^{-t}$ in statistical distance of independent U_{2t^2} and U_t. Applying G to the second component and using its security proves that $Z \| G(s)$ is pseudorandom. The full proof appears in Proposition E.1. □

6 Consequences

6.1 Sequential Sampling and Cryptography

State elimination rules out generating several cryptographic distributions with logarithmic private state, even when the entire output prefix remains available. The following general consequence permits visible information W, such as a public key.

Corollary 6.1 (General consequence for samplers with logarithmic state)**.** *Let (W, Y) be jointly distributed and indexed by n. Assume:*

1. *W has polynomial length, and Y consists of $\mathsf{poly}(n)$ items over a $\mathsf{poly}(n)$-size alphabet.*
2. *Y has a noticeable nonuniform next-block pseudoentropy gap given W.*

Then no sequential sampler C with $\mathsf{poly}(n)$-size circuits and $O(\log n)$ private bits satisfies

$$D_{\mathsf{KL}}\left((W, Y) \parallel (W, C(W))\right) = \mathsf{negl}(n),$$

even given W and the entire visible prefix at every step.

The full proof appears in Corollary G.1. We now state the resulting cryptographic limitations.

Corollary 6.2 (Pseudorandom generation)**.** *Let $G : \{0,1\}^{k(n)} \to \{0,1\}^n$ be a pseudorandom generator secure against nonuniform polynomial-size circuits. No sequential bit sampler C with $\mathsf{poly}(n)$-size circuits and $O(\log n)$ private bits satisfies*

$$D_{\mathsf{KL}}\left(G(U_{k(n)}) \parallel C(1^n)\right) = \mathsf{negl}(n).$$

A *public-key encryption scheme* generates a public key pk and a secret key sk. Anyone can encrypt a message using pk; decryption using sk recovers it. *IND-CPA security* requires indistinguishable encryptions of any two equal-length messages chosen efficiently after seeing pk. *Public-key steganography* additionally requires encodings to be indistinguishable from a prescribed cover distribution, such as ordinary text. The following results concern this public-key setting, without a pre-shared secret between sender and receiver.

Corollary 6.3 (Public-key encryption)**.** *Let a polynomial-time public-key encryption scheme have security parameter n. Assume:*

1. *IND-CPA security against nonuniform polynomial-size circuits and negligible correctness error.*
2. *Ciphertexts consist of a fixed polynomially bounded number of $O(\log n)$-bit tokens.*

For any two distinct equal-length messages of polynomial length, their ciphertext distributions cannot both be sampled with negligible KL-divergence by sequential samplers with $\mathsf{poly}(n)$-size circuits and $O(\log n)$ private bits. The public key, message, and visible prefix are available at every step; divergence is measured jointly with the public key.

Corollary 6.4 (Public-key steganography)**.** *Let a public-key steganography scheme run in polynomial time, with security parameter n. Fix two distinct messages of polynomial length. Assume:*

1. *Negligible decoding error; both encodings consist of the same fixed polynomially bounded number of $O(\log n)$-bit tokens.*

2. *Both encodings are computationally indistinguishable from the same cover distribution, jointly with the public key, against nonuniform polynomial-size circuits.*

The two encoding distributions cannot both be sampled with negligible KL-divergence by sequential samplers with $\mathsf{poly}(n)$*-size circuits and* $O(\log n)$ *private bits. The public key, message, and visible prefix are available at every step; divergence is measured jointly with the public key.*

Proof overview. For encryption and steganography, the two encodings $P_b = (\mathsf{pk}, Y^b)$ are computationally indistinguishable but have statistical distance $1 - \mathsf{negl}(n)$. Indeed, adjoining the secret key preserves statistical distance, since it is independent of the encoding given the public key, and decoding distinguishes the messages. For a PRG, use $P_0 = G(U_{k(n)})$ and $P_1 = U_n$ instead: the support-size bound gives $\Delta(P_0, P_1) \geq 1 - 2^{k(n)-n} \geq 1/2$. Let N be the number of tokens and u the alphabet size.

1. Suppose both distributions admit the claimed samplers. Put $\alpha = 1/(100N)$. By Theorem 4.2 and Pinsker's inequality, obtain polynomial-size stateless samplers D^b whose distributions R_b satisfy $\Delta(P_b, R_b) \leq \alpha$. Hence $\Delta(R_0, R_1) \geq 1/3$ for sufficiently large n. For the PRG comparison, the sampler for P_1 is simply uniform.
2. A hybrid replacing one transition at a time gives an index i at which D_i^0 and D_i^1, on a common prefix from R_0, have average statistical distance at least $1/(3N)$. The prefix includes the public key when present.
3. Estimate each transition's token probabilities using $M = \lceil 576uN^2 \rceil$ fresh samples, and accept when the observed token has larger estimated probability under D_i^0. The sum of the expected ℓ_1 estimation errors is at most $2\sqrt{u/M} \leq 1/(12N)$, so this test has advantage at least $1/(4N)$. Yet replacing the prefixes through P_0, P_1 and using their indistinguishability bounds its advantage by $4\alpha + \mathsf{negl}(n)$, a contradiction.

The index i is nonuniform advice; all sampling and tests have polynomial size. Detailed proofs appear with Corollaries G.2 to G.4. □

6.2 A Case For Insertional Sampling

Stern et al. (2019) introduce an insertion framework and implement it with the *Insertion Transformer*. Given an input sequence x and the current partial output w, called the *canvas*, the model assigns a joint distribution $p(a, j \mid x, w)$ to the next token a and its insertion position j. The decoder's hidden states are recomputed after each insertion, so earlier computation leaves no private state beyond the canvas (Stern et al., 2019, Section 4.5).

To connect this framework to ours, fix the input x and the output length, and consider sampling one insertion per step with fresh coins. This is a sampling version of their framework; their decoding procedures choose maximum-probability actions, either serially or in parallel (Stern et al., 2019, Section 5). Allowing arbitrary randomized circuits for these transitions gives exactly our stateless insertional samplers, up to polynomial overhead (Definition D.1). Indeed, a token and position determine the next canvas; conversely, a one-token extension determines a valid insertion by locating a token whose deletion recovers the old canvas. Restricting every insertion to the rightmost position gives stateless sequential sampling.

Our separation now gives a computational motivation for insertion-based models. Proposition F.1 shows that efficient insertional generation can exactly sample distributions over binary strings that efficient left-to-right generation cannot sample even with small KL error. The advantage already appears with one insertion per step: the generator can produce information needed for later decisions before filling in earlier output positions. Thus choosing where to write can expand what is efficiently generatable without adding private memory or requiring parallel decoding. Indeed, as we have shown in Proposition 5.5, while PRGs are hard to sample for sequential samplers, even with $O(\log n)$ bits of state, this is not necessarily the case for insertional samplers.

This conclusion concerns the generation rule with general efficient transitions; realizing the separating construction in a particular Transformer architecture requires a separate argument. We leave models allowing deletions or substitutions to future work.

AI USE STATEMENT

The core research ideas are due to the authors. We used large language models, specifically ChatGPT 6 Astra, to assist with writing and revising the manuscript, including proofs, and to help check mathematical arguments for gaps and errors. The authors take full responsibility for the final content of the paper, including its claims, proofs, and references.

REFERENCES

Rohit Agrawal, Yi-Hsiu Chen, Thibaut Horel, and Salil P. Vadhan. Unifying computational entropies via kullback-leibler divergence. In Alexandra Boldyreva and Daniele Micciancio (eds.), *CRYPTO 2019, Part II*, volume 11693 of *LNCS*, pp. 831–858. Springer, Cham, August 2019. doi: 10.1007/978-3-030-26951-7_28.

Kenneth E. Batcher. Sorting networks and their applications. In *Proceedings of the Spring Joint Computer Conference*, AFIPS '68 (Spring), pp. 307–314. ACM, 1968. doi: 10.1145/1468075.1468121.

Clément L Canonne, Ziteng Sun, and Ananda Theertha Suresh. Concentration bounds for discrete distribution estimation in kl divergence. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 2093–2098. IEEE, 2023.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Lance Fortnow. How does machine learning manage complexity?, 2026. URL https://arxiv.org/abs/2604.07233.

Iftach Haitner, Omer Reingold, Salil P. Vadhan, and Hoeteck Wee. Inaccessible entropy. In Michael Mitzenmacher (ed.), *41st ACM STOC*, pp. 611–620. ACM Press, May / June 2009. doi: 10.1145/1536414.1536497.

Russell Impagliazzo and Leonid A. Levin. No better ways to generate hard NP instances than picking uniformly at random. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, volume 2, pp. 812–821. IEEE, 1990. doi: 10.1109/FSCS.1990.89604.

Haozhe Jiang, Nika Haghtalab, and Lijie Chen. Diffusion language models are provably optimal parallel samplers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2f992baabfee60ab6add5e4f63b414b1-Abstract-Conference.html.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30749–30768. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25ah.html.

Leonid A. Levin. Average case complete problems. *SIAM Journal on Computing*, 15(1):285–286, 1986. doi: 10.1137/0215020. URL https://doi.org/10.1137/0215020.

Chu-Cheng Lin, Aaron Jaech, Xin Li, Matthew R. Gormley, and Jason Eisner. Limitations of autoregressive models and their alternatives. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5147–5173. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.405. URL https://aclanthology.org/2021.naacl-main.405/.

Jaouad Mourtada and Stéphane Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research*, 23(31):1–49, 2022. URL https://jmlr.org/papers/v23/20-782.html.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. Insertion transformer: Flexible sequence generation via insertion operations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5976–5985. PMLR, 2019. URL http://proceedings.mlr.press/v97/stern19a.html.

Salil P. Vadhan and Colin Jia Zheng. Characterizing pseudoentropy and simplifying pseudorandom generator constructions. In Howard J. Karloff and Toniann Pitassi (eds.), *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pp. 817–836. ACM, 2012. doi: 10.1145/2213977.2214051. URL https://doi.org/10.1145/2213977.2214051.

Gabriel Wu. Theoretical limitations of autoregressive models, May 2023. URL https://gabrieldwu.com/limitations-autoregressive. Blog post, May 28.

Chenxiao Yang, Cai Zhou, David Wipf, and Zhiyuan Li. On powerful ways to generate: Autoregression, diffusion, and beyond. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/bf5b94743490c1488027ee403188b4c5-Abstract-Conference.html.

A Additional Preliminaries

For $n \in \mathbb{N}$, define the set $[n] = \{1, \ldots, n\}$. For any set X, we define $X^0 := \{\emptyset\}$ where $\emptyset$ denotes the empty string. We say that a string x is a subsequence of y if x can be produced from y by deletions.

For any two random variables or two distributions X, Y, we will use $\Delta(X, Y) := \frac{1}{2} \cdot \sum_{a \in \mathsf{supp}(X) \cup \mathsf{supp}(Y)} |\Pr[X = a] - \Pr[Y = a]| \in [0, 1]$ to denote the total variation distance between X and Y. Our default logarithm function $\log$ is in base two. If X, Y are random variables where Y has support $\mathcal{Y}$, then for any $y \in \mathcal{Y}$ we use the notation $X\,|_{Y=y}$ to denote the conditional distribution of X conditioned on the event that $Y = y$.

A function f in parameter $\lambda \in \mathbb{N}$ is noticeable if there exists a polynomial $p(\lambda)$ such that for all sufficiently large λ, $f(\lambda) > 1/p(\lambda)$. For a parameter $\lambda \in \mathbb{N}$, for any two functions $\alpha = \alpha(\lambda)$ and $\beta = \beta(\lambda)$, we say that β is noticeably larger α if there exists a noticeable function $\gamma = \gamma(\lambda)$ such that $\beta > \alpha + \gamma$. A function f in parameter $\lambda \in \mathbb{N}$ is negligible if for all polynomials $p(\lambda)$, for sufficiently large enough λ, we have $f(\lambda) < 1/p(\lambda)$.

A.1 Information Theory

We now recall relevant definitions and related facts from information theory.

Definition A.1 (Shannon Entropy)**.** *For a discrete random variable X with finite support, we define the Shannon entropy of X to be $H(X) := -\mathbb{E}_{x \sim X}[\log_2 \Pr(X = x)]$.*

Definition A.2 (Binary Entropy)**.** *For $p \in [0, 1]$, define the binary entropy function $h(p) = -p \log_2 p - (1 - p) \log_2 (1 - p)$.*

Definition A.3 (KL Divergence)**.** *Let P, Q be distributions over a finite set Ω. If P is absolutely continuous with respect to Q, meaning that $P(a) > 0$ implies $Q(a) > 0$, then the KL divergence from P to Q is*

$$D_{\mathrm{KL}}(P \| Q) := \sum_{a \in \Omega} P(a) \log \frac{P(a)}{Q(a)}.$$

If P is not absolutely continuous with respect to Q, we define $D_{\mathrm{KL}}(P \| Q) = +\infty$.

Fact A.4 (Nonnegativity of KL Divergence; see Theorem 2.6.3 in Cover & Thomas (2006))**.** *For any two distributions P, Q over a finite set Ω, $D_{\mathrm{KL}}(P \| Q) \geq 0$.*

Fact A.5 (Pinsker's inequality; Lemma 11.6.1 in Cover & Thomas (2006))**.** *For finite distributions P, Q,*

$$\Delta(P, Q) \leq \sqrt{(\ln 2) D_{\mathsf{KL}}(P \parallel Q)/2}.$$

Definition A.6 (Markov Chain; see Chapter 4 in Cover & Thomas (2006)). *A sequence of random variables $Z_1, \ldots, Z_n$ over finite sets is a Markov chain, denoted $Z_1 - Z_2 - \cdots - Z_n$ if for every $i \in \{2, \ldots, n-1\}$,*

$$\Pr[Z_{i+1} = z_{i+1} \mid Z_1 = z_1, \ldots, Z_i = z_i] = \Pr[Z_{i+1} = z_{i+1} \mid Z_i = z_i]$$

whenever the conditioning event has positive probability.

Equivalently, for every $i \in \{2, \ldots, n-1\}$, Z_{i+1} is conditionally independent of $(Z_1, \ldots, Z_{i-1})$ given Z_i.

The KL chain rule compares two joint distributions by first comparing one variable, then comparing the other conditioned on the first.

Fact A.7 (KL chain rule; Theorem 2.5.3 in Cover & Thomas (2006)). *Let P_{XY} and Q_{XY} be distributions over the same finite set $\mathcal{X} \times \mathcal{Y}$. Then*

$$D_{\mathsf{KL}}\left(P_{XY} \,\|\, Q_{XY}\right) = D_{\mathsf{KL}}\left(P_X \,\|\, Q_X\right) + \mathop{\mathbb{E}}_{x \sim P_X} D_{\mathsf{KL}}\left(P_{Y|X=x} \,\|\, Q_{Y|X=x}\right).$$

Conditional distributions on events of probability zero may be chosen arbitrarily, and the expectation sums only over x with $P_X(x) > 0$. The identity allows infinite divergence.

Iterating the chain rule gives a sum over steps, each conditioned on the full preceding history. For Markov chains, the conditioning reduces to the preceding variable.

Fact A.8 (KL chain rule for Markov chains). *Let $n \geq 1$ and let $X_0, \ldots, X_n$ and $Y_0, \ldots, Y_n$ be Markov chains, where for each i the variables X_i, Y_i take values in the same finite set, which may depend on i. If they have the same initial distribution, $X_0 \equiv Y_0$, then*

$$D_{\mathsf{KL}}\left((X_0, \ldots, X_n) \,\|\, (Y_0, \ldots, Y_n)\right) = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{x \sim X_{i-1}} D_{\mathsf{KL}}\left(X_i \,|_{X_{i-1}=x} \,\|\, Y_i \,|_{Y_{i-1}=x}\right). \tag{1}$$

Both transitions are conditioned on the same value x, sampled from X_{i-1}, with the same conventions for zero-probability events as above.

Proof. Applying Fact A.7 to the first coordinate gives

$$\begin{aligned} &D_{\mathsf{KL}}\left((X_0, \ldots, X_n) \,\|\, (Y_0, \ldots, Y_n)\right) \\ &\quad = D_{\mathsf{KL}}\left(X_0 \,\|\, Y_0\right) + \mathop{\mathbb{E}}_{x_0 \sim X_0} D_{\mathsf{KL}}\left((X_1, \ldots, X_n) \,|_{X_0=x_0} \,\|\, (Y_1, \ldots, Y_n) \,|_{Y_0=x_0}\right). \end{aligned}$$

Here $D_{\mathsf{KL}}\left(X_0 \,\|\, Y_0\right) = 0$ because $X_0 \equiv Y_0$. For $n = 1$, this already gives the claim. For $i \in [n]$, write $X_{<i} = (X_0, \ldots, X_{i-1})$ and $Y_{<i} = (Y_0, \ldots, Y_{i-1})$. For $n \geq 2$, applying the same fact inside the expectation separates the next coordinate:

$$\begin{aligned} &D_{\mathsf{KL}}\left((X_0, \ldots, X_n) \,\|\, (Y_0, \ldots, Y_n)\right) \\ &\quad = D_{\mathsf{KL}}\left(X_0 \,\|\, Y_0\right) + \mathop{\mathbb{E}}_{x_0 \sim X_0} D_{\mathsf{KL}}\left(X_1 \,|_{X_0=x_0} \,\|\, Y_1 \,|_{Y_0=x_0}\right) \\ &\qquad + \mathop{\mathbb{E}}_{(x_0,x_1) \sim (X_0,X_1)} D_{\mathsf{KL}}\left((X_2, \ldots, X_n) \,|_{X_0=x_0,\, X_1=x_1} \,\|\, (Y_2, \ldots, Y_n) \,|_{Y_0=x_0,\, Y_1=x_1}\right) \\ &\quad = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{x_{<i} \sim X_{<i}} D_{\mathsf{KL}}\left(X_i \,|_{X_{<i}=x_{<i}} \,\|\, Y_i \,|_{Y_{<i}=x_{<i}}\right). \end{aligned}$$

The Markov property gives

$$\begin{aligned} X_i \,|_{X_{<i}=x_{<i}} &\equiv X_i \,|_{X_{i-1}=x_{i-1}}, \\ Y_i \,|_{Y_{<i}=x_{<i}} &\equiv Y_i \,|_{Y_{i-1}=x_{i-1}}. \end{aligned}$$

Consequently, tuples $x_{<i}$ with the same last value $x_{i-1} = x$ give the same KL divergence. Their weights in the expectation add up to $\Pr[X_{i-1} = x]$:

$$\begin{aligned} &\mathop{\mathbb{E}}_{x_{<i} \sim X_{<i}} D_{\mathsf{KL}}\left(X_i \,|_{X_{<i}=x_{<i}} \,\|\, Y_i \,|_{Y_{<i}=x_{<i}}\right) \\ &\quad = \sum_{x} \left(\sum_{x_{<i}:\, x_{i-1}=x} \Pr[X_{<i} = x_{<i}] \right) D_{\mathsf{KL}}\left(X_i \,|_{X_{i-1}=x} \,\|\, Y_i \,|_{Y_{i-1}=x}\right) \\ &\quad = \mathop{\mathbb{E}}_{x \sim X_{i-1}} D_{\mathsf{KL}}\left(X_i \,|_{X_{i-1}=x} \,\|\, Y_i \,|_{Y_{i-1}=x}\right). \end{aligned}$$

Substituting these equalities into the expansion proves (1). □

The KL inequality below follows from the log-sum inequality (Cover & Thomas, 2006, Theorem 2.7.1).

Fact A.9 (Data Processing Inequality). *Let P_X, Q_X be distributions on a finite set $\mathcal{X}$. Let P_Y, Q_Y be obtained by applying the same randomized map from $\mathcal{X}$ to a finite set $\mathcal{Y}$. Then $D_{\mathsf{KL}}(P_Y \parallel Q_Y) \le D_{\mathsf{KL}}(P_X \parallel Q_X)$.*

Definition A.10 (Laplace Estimator (Canonne et al., 2023)). *Let P be a distribution over a finite domain Ω of size $|\Omega| = K$. For $N > 0$, let $\bar{x} = x^{(1)}, \dots x^{(N)}$ be i.i.d. samples according to P. The Laplace estimator is defined (w.r.t. $\bar{x}$) for any $\omega \in \Omega$ as*

$$\hat{P}_N(\omega) = \frac{1}{N+K}\left(1 + \sum_{j=1}^{N} \mathbb{1}(x^{(j)} = \omega)\right) \tag{2}$$

The following claim extends the standard expected KL bound for the Laplace estimator (the case $Q = P$; see Mourtada & Gaïffas, 2022, Proposition 4) to an arbitrary target distribution Q.

Claim A.11. *Let P, Q be distributions over a finite domain Ω of size $K \ge 1$, and let $N \ge 1$ be an integer. Let $\hat{P}_N$ be the Laplace estimate from N independent samples of P. Then*

$$\mathbb{E}\, D_{\mathsf{KL}}\left(Q \parallel \hat{P}_N\right) \le D_{\mathsf{KL}}(Q \parallel P) + \log\frac{N+K}{N+1} \le D_{\mathsf{KL}}(Q \parallel P) + \frac{(K-1)\log e}{N+1},$$

where the expectation is over the samples and logarithms have base two.

Proof. If Q is not absolutely continuous with respect to P, the right-hand side is infinite and the claim holds. Otherwise, fix $x \in \mathsf{supp}(Q)$ and put $p = P(x) > 0$. Its sample count V_x has distribution $\mathrm{Bin}(N, p)$, and $\hat{P}_N(x) = (V_x + 1)/(N + K)$. Hence

$$\begin{aligned}\mathbb{E}\left[\frac{P(x)}{\hat{P}_N(x)}\right] &= p(N+K)\,\mathbb{E}\left[\frac{1}{V_x+1}\right] \\ &= \frac{N+K}{N+1}\left(1-(1-p)^{N+1}\right) \le \frac{N+K}{N+1}.\end{aligned}$$

The second equality follows by expanding the binomial expectation and using $\binom{N}{v}/(v+1) = \binom{N+1}{v+1}/(N+1)$. Expanding KL divergence and applying Jensen's inequality now gives

$$\begin{aligned}\mathbb{E}\, D_{\mathsf{KL}}\left(Q \parallel \hat{P}_N\right) &= D_{\mathsf{KL}}(Q \parallel P) + \sum_{x \in \mathsf{supp}(Q)} Q(x)\,\mathbb{E}\log\frac{P(x)}{\hat{P}_N(x)} \\ &\le D_{\mathsf{KL}}(Q \parallel P) + \sum_{x \in \mathsf{supp}(Q)} Q(x)\log\mathbb{E}\left[\frac{P(x)}{\hat{P}_N(x)}\right] \\ &\le D_{\mathsf{KL}}(Q \parallel P) + \log\frac{N+K}{N+1}.\end{aligned}$$

Finally, $\log(1+y) \le y\log e$ for $y \ge 0$ gives the second bound. □

A.2 One-Way Functions

We use the length-preserving form of one-way functions, with security against nonuniform circuits.

Definition A.12 (One-Way Functions against Circuits). *Let $f = (f_\lambda)_{\lambda \ge 1}$ be a function family with $f_\lambda : \{0,1\}^\lambda \to \{0,1\}^\lambda$, where $f_\lambda(x)$ is computable from $(1^\lambda, x)$ in deterministic polynomial time. For functions $T : \mathbb{N} \to [1,\infty)$ and $\varepsilon : \mathbb{N} \to [0,1]$, we say that f is (T, ε)-one-way against circuits if, for all sufficiently large λ, every randomized circuit A_λ of size at most $T(\lambda)$ satisfies*

$$\Pr_{X \leftarrow U_\lambda, R}\left[A_\lambda(f_\lambda(X); R) \in f_\lambda^{-1}(f_\lambda(X))\right] \le \varepsilon(\lambda).$$

Here R denotes independent uniform random coins for A_λ. The circuit may depend arbitrarily on λ, and success means finding any preimage; an output $\perp$ counts as failure.

We call f a one-way function against polynomial-size circuits *if it is $(\lambda^c, \lambda^{-c})$-one-way for every constant $c > 0$. For a size bound T, we call f T-*secure *if it is $(T, 1/T)$-one-way.*

Proposition F.1 uses one-wayness against polynomial-size circuits. Theorem F.2 assumes the stronger subexponential security $T(\lambda) = 2^{\gamma\lambda^{\alpha}}$ for constants $\gamma > 0$ and $1/2 < \alpha < 1$.

A.3 PSEUDOENTROPY

Definition A.13 (Pseudoentropy, nonuniform setting (Vadhan & Zheng, 2012))**.** *Let X be a random variable over a finite set. We say that X has (T, ε) nonuniform pseudoentropy at least k if there exists a random variable Y over the same set with*

1. *$H(Y) \geq k$ and*
2. *X and Y are (T, ε) nonuniformly indistinguishable.*

Here indistinguishability means that every Boolean circuit of size at most T has distinguishing advantage at most ε in absolute value.

If $X = X(n)$, we say that X has nonuniform pseudoentropy at least $k = k(n)$ if, for every constant $c > 0$, it has (n^c, n^{-c}) nonuniform pseudoentropy at least $k(n) - n^{-c}$ for all sufficiently large n.

Definition A.14 (Conditional pseudoentropy, nonuniform setting (Vadhan & Zheng, 2012))**.** *Let (X, B) take values in the finite set $\mathcal{X} \times \mathcal{B}$. We say that B has (T, ε) conditional pseudoentropy at least k given X if there exists a random variable Z over $\mathcal{B}$, jointly distributed with X, such that*

1. *$H(Z \mid X) \geq k$ and*
2. *(X, B) and (X, Z) are (T, ε) nonuniformly indistinguishable.*

The conditional pseudoentropy is the supremum of such k. The (T, ε) conditional pseudoentropy gap of B given X is this supremum minus $H(B \mid X)$. The asymptotic definition uses the same quantifiers and entropy slack as Definition A.13.

The maximum exists: the conditional distributions of Z form a compact set, the indistinguishability constraints are closed, and conditional entropy is continuous. Taking $Z = B$ shows that the gap is nonnegative.

Definition A.15 (Next-block pseudoentropy (Vadhan & Zheng, 2012, Definition 2.12))**.** *Let $n, \ell \geq 1$, let $Y = (Y_1, \ldots, Y_n)$ take values in $(\{0,1\}^{\ell})^n$, and let I be uniform on $[n]$, independently of Y. We say that Y has (t, ϵ) next-block pseudoentropy at least k if there exists a random variable Z over $\{0,1\}^{\ell}$, jointly distributed with (I, Y), such that*

1. *$H(Z \mid Y_{<I}) \geq k/n$ and*
2. *$(Y_{<I}, Y_I)$ and $(Y_{<I}, Z)$ are (t, ϵ) nonuniformly indistinguishable.*

For $\ell = 1$, we also call this next-bit pseudoentropy.

For families indexed by n, next-block pseudoentropy at least $k(n)$ means that, for every $c > 0$ and all sufficiently large n, the (n^c, n^{-c}) next-block pseudoentropy is at least $k(n) - n^{-c}$. This is the nonuniform asymptotic convention; polynomial rescaling absorbs the number of blocks when comparing it with the conditional definition.

Definition A.16 (KL-hardness for conditional sampling (Vadhan & Zheng, 2012))**.** *Let (X, B) take values in $\{0,1\}^m \times [q]$. We say that B is (t, δ) KL-hard to sample given X if every randomized circuit $S : \{0,1\}^m \to [q]$ of size at most t satisfies $D_{\mathsf{KL}}((X, B) \,\|\, (X, S(X))) > \delta$.*

We use the following two consequences of the Vadhan–Zheng characterization, stated directly in terms of conditional samplers. In the first part, the distinguishing error ϵ and the additive KL-divergence loss η are separate parameters.

Theorem A.17 (Characterizing conditional pseudoentropy, nonuniform setting)**.** *There exist a universal constant $c_1 \in (0, 1]$ and a polynomial p_0 with nonnegative coefficients and constant term at least one such that the following holds. Let (X, B) take values in $\{0,1\}^m \times [q]$, with $m \geq 0$ and $q \geq 2$, and let $t \geq 1$, $\delta, g \geq 0$, and $0 < \epsilon, \eta, \sigma < 1$.*

1. *If B has (t, ϵ) conditional pseudoentropy gap at least δ given X, then every randomized circuit $S : \{0,1\}^m \to [q]$ of size at most*

$$\frac{\min\{t^{c_1},\, \sigma/\epsilon\}}{p_0(m, q, 1/\eta, \log(1/\sigma))}$$

satisfies $D_{\mathsf{KL}}\left((X, B) \parallel (X, S(X))\right) > \delta - \sigma - \eta$.

2. *If the (t, η) conditional pseudoentropy gap of B given X is at most g, then there exists a randomized circuit $S : \{0,1\}^m \to [q]$ of size at most $p_0(t, m, q, 1/\eta)$ such that*

$$D_{\mathsf{KL}}\left((X, B) \parallel (X, S(X))\right) \le g + 2\eta.$$

Proof. The first part combines (Vadhan & Zheng, 2012, Theorem 3.8 and Lemma 3.6), with respective KL-divergence losses σ and η. Choosing c_1 sufficiently small and p_0 sufficiently large gives the stated circuit sizes. For $m = 0$, add a fixed conditioning bit; for $\delta = 0$, the first part follows from nonnegativity of KL-divergence.

For the second part, suppose that every sampler of size at most $p_0(t, m, q, 1/\eta)$ has KL-divergence greater than $g+2\eta$. For sufficiently large p_0, the same results give a (t, η) conditional pseudoentropy gap at least $g + \eta$, contradicting its upper bound g. One polynomial p_0 can be chosen large enough for both parts, absorbing constant shifts of its arguments. □

The running time of the resulting sampler is polynomial in the inverse accuracy. Consequently, the finite-parameter statement supplies a family of polynomial-size samplers for inverse-polynomial accuracies.

B Characterization of Sequential Sampling

We apply Theorem A.17 with the visible prefix as the conditioning variable and the next block as the variable to be sampled. Choosing the block uniformly gives a characterization of stateless sequential sampling. The dependence on the alphabet size 2^ℓ is polynomial, contributing a polynomial factor in n when $\ell = O(\log n)$.

Theorem B.1 (Characterization of stateless sequential sampling)**.** *There exist a universal constant $c_0 \in (0, 1]$ and a polynomial p with nonnegative coefficients and constant term at least one such that the following holds. Let $n, \ell \ge 1$, let $Y = (Y_1, \ldots, Y_n)$ take values in $(\{0,1\}^\ell)^n$, and let $t \ge 1$, $\delta \ge 0$, and $0 < \epsilon, \eta, \sigma < 1$.*

1. *If Y has (t, ϵ) next-block pseudoentropy at least $H(Y) + \delta$, then, for every $B \ge 1$ satisfying*

$$B \le \frac{\min\{t^{c_0},\, \sigma/(n\epsilon)\}}{p(n, 2^\ell, 1/\eta, \log(n/\sigma))},$$

every stateless (ℓ, B)-sequential sampler C satisfies $D_{\mathsf{KL}}\left(Y \parallel C(1^n)\right) > \delta - \sigma - n\eta$.

2. *If $\delta \ge \epsilon$ and the (t, ϵ) next-block pseudoentropy of Y is at most $H(Y) + \delta - \epsilon$, then there exists a stateless (ℓ, B)-sequential sampler C with*

$$B \le p(t, n, 2^\ell, 1/\epsilon), \qquad D_{\mathsf{KL}}\left(Y \parallel C(1^n)\right) \le \delta.$$

We prove the two directions separately. Throughout, I is uniform on $[n]$ and independent of Y, and $X = (I, Y_{<I})$. Write $m = O(n\ell)$ for the input length of X and $q = 2^\ell$ for the alphabet size.

Lemma B.2 (Next-block pseudoentropy implies conditional sampling hardness)**.** *Under the hypotheses and size condition of the first part of Theorem B.1, every stateless (ℓ, B)-sequential sampler C satisfies*

$$\mathop{\mathbb{E}}_{i \leftarrow [n]} D_{\mathsf{KL}}\left((Y_{<i}, Y_i) \parallel (Y_{<i}, C_i(Y_{<i}))\right) > \frac{\delta - \sigma}{n} - \eta.$$

Proof. By Definition A.15, there is a random variable Z jointly distributed with (I, Y) such that (X, Z) and (X, Y_I) are (t, ϵ)-indistinguishable and

$$H(Z \mid Y_{<I}) = \frac{1}{n}\sum_{i=1}^{n} H(Z \mid Y_{<i}, I = i) \geq \frac{H(Y) + \delta}{n}.$$

Here $Y_{<I}$ determines I. Since $H(Y_I \mid X) = H(Y)/n$, the (t, ϵ) conditional pseudoentropy gap of Y_I given X is at least δ/n. This is an average over the index and does not give the same gap for each fixed i. Fix a stateless sampler C. To apply Theorem A.17, we therefore need a sampler that, given $(I, Y_{<I})$, samples exactly $C_I(Y_{<I})$. The loss η in that theorem comes from approximating a sampler's output probabilities to obtain the predictor used in its proof.

We now construct this sampler, which we call S. It contains one copy of each C_j and, on input $(i, y_{<i})$, selects C_i to produce the output $C_i(y_{<i})$. The size increases, but remains sufficiently small for the first part of Theorem A.17 to apply. Namely, we implement this selection by ANDing each output bit of C_j with the comparison $i = j$, and then ORing these bits over $j \in [n]$. This sets every unselected output to zero, leaving only the output of C_i.

Each C_j contributes at most B gates, with $O(n\ell)$ more for its index comparison and output masking, giving $B + O(n\ell)$ gates per copy. With n copies and $O(n\ell)$ gates for the final ORs, the size of S is $O(n(B + n\ell)) = O(n^2 qB)$. Take $c_0 = c_1$. Since $m = O(nq)$ and p_0 has nonnegative coefficients, choose the fixed polynomial p sufficiently large to absorb the $O(n^2 q)$ overhead and the dependence of p_0 on m. The assumed bound on B then makes the size of S at most

$$\frac{\min\{t^{c_1},\, \sigma/(n\epsilon)\}}{p_0(m, q, 1/\eta, \log(n/\sigma))}.$$

Therefore, the first part of Theorem A.17, with gap δ/n, loss σ/n, and the same t, ϵ, and η, gives

$$\begin{aligned} D_{\mathsf{KL}}\left((X, Y_I) \parallel (X, S(X))\right) &= \frac{1}{n}\sum_{i=1}^{n} D_{\mathsf{KL}}\left((Y_{<i}, Y_i) \parallel (Y_{<i}, C_i(Y_{<i}))\right) \\ &> \frac{\delta - \sigma}{n} - \eta. \end{aligned} \tag{3}$$

The equality conditions on the common uniform index I. □

Lemma B.3 (Small next-block pseudoentropy gap gives samplers)**.** *Under the hypotheses of the second part of Theorem B.1, there is a stateless (ℓ, B)-sequential sampler C, with $B \leq p(t, n, 2^\ell, 1/\epsilon)$, such that*

$$\mathop{\mathbb{E}}_{i \leftarrow [n]} D_{\mathsf{KL}}\left((Y_{<i}, Y_i) \parallel (Y_{<i}, C_i(Y_{<i}))\right) \leq \frac{\delta}{n}.$$

Proof. By assumption, the (t, ϵ) next-block pseudoentropy of Y is at most $H(Y) + \delta - \epsilon$. We use this bound in the second part of Theorem A.17 to obtain a sampler S that receives $(i, y_{<i})$ and samples Y_I given X up to KL-divergence δ/n. Fixing its index input to i gives the circuit C_i: on input $y_{<i}$, it outputs $S(i, y_{<i})$.

By Definition A.15 and $H(Y_I \mid X) = H(Y)/n$, the (t, ϵ) conditional pseudoentropy gap of Y_I given X is at most $(\delta - \epsilon)/n$. The target KL-divergence δ/n leaves ϵ/n for the theorem's additive loss 2η. We therefore set $\eta' = \epsilon/(2n)$. Since $\eta' \leq \epsilon$, every distribution satisfying (t, η')-indistinguishability also satisfies (t, ϵ)-indistinguishability, so the same upper bound on the gap holds at error η'. Since $\delta \geq \epsilon$ and $0 < \eta' < 1$, we can therefore apply that theorem with $g = (\delta - \epsilon)/n$ and η' in place of η. It gives a randomized circuit S of size at most $p_0(t, m, q, 2n/\epsilon)$ such that

$$D_{\mathsf{KL}}\left((X, Y_I) \parallel (X, S(X))\right) \leq \frac{\delta - \epsilon}{n} + 2\eta' = \frac{\delta}{n}.$$

For each $i \in [n]$, define $C_i(y_{<i}) = S(i, y_{<i})$ by hardwiring the index input to i. This does not increase the circuit size. Using fresh independent coins at every step, the circuits C_i form a stateless sequential sampler. Since $m = O(nq)$, their size is bounded by a fixed polynomial in $t, n, q, 1/\epsilon$. Enlarge the polynomial p from the preceding lemma to dominate this bound, retaining nonnegative

coefficients and constant term at least one. This is compatible with that lemma because enlarging p only strengthens its size hypothesis. Thus $B \le p(t, n, 2^\ell, 1/\epsilon)$.

Finally, the two distributions in the KL-divergence have the same uniform index I. Conditioning on this index gives

$$\mathop{\mathbb{E}}_{i \leftarrow [n]} D_{\mathsf{KL}}\left((Y_{<i}, Y_i) \parallel (Y_{<i}, C_i(Y_{<i}))\right) = D_{\mathsf{KL}}\left((X, Y_I) \parallel (X, S(X))\right) \le \frac{\delta}{n}.$$

□

Proof of Theorem B.1. Since the sampler is stateless and uses fresh coins at each step, $\Pr[C(1^n) = y] = \prod_i \Pr[C_i(y_{<i}) = y_i]$. Together with the analogous factorization of $\Pr[Y = y]$, this gives

$$\begin{aligned} D_{\mathsf{KL}}\left(Y \parallel C(1^n)\right) &= \sum_{i=1}^{n} D_{\mathsf{KL}}\left((Y_{<i}, Y_i) \parallel (Y_{<i}, C_i(Y_{<i}))\right) \\ &= n \mathop{\mathbb{E}}_{i \leftarrow [n]} D_{\mathsf{KL}}\left((Y_{<i}, Y_i) \parallel (Y_{<i}, C_i(Y_{<i}))\right). \end{aligned}$$

Indeed, when the divergence is finite, its definition gives

$$D_{\mathsf{KL}}\left(Y \parallel C(1^n)\right) = \sum_{y \in \mathsf{supp}(Y)} \Pr[Y = y] \sum_{i=1}^{n} \log \frac{\Pr[Y_i = y_i \mid Y_{<i} = y_{<i}]}{\Pr[C_i(y_{<i}) = y_i]}.$$

For each i, marginalizing over $y_{>i}$ identifies its contribution as $D_{\mathsf{KL}}\left((Y_{<i}, Y_i) \parallel (Y_{<i}, C_i(Y_{<i}))\right)$. If some $y \in \mathsf{supp}(Y)$ has zero probability under C, at least one of its next-block probabilities is zero; the corresponding summand and the total KL-divergence are both infinite. The first conclusion now follows from Lemma B.2, and the second from Lemma B.3. □

C Stateful Sequential Sampling

We now examine the setting in which circuit C_i can pass k bits of state to circuit C_{i+1}. Stateful sequential sampling captures this model and we show that state can be removed by blowing up the size of the circuits C_i. Namely, if there is a distribution D can be sampled by a stateful sequential sampler, then it can be sampled by a stateless sequential sampler with a 2^k blow-up in circuit size up to polynomial factors. This tradeoff is tight in the worst-case (for estimating an unknown state distribution from samples) as we will show, i.e. there are distributions for which this blow-up is necessary.

C.1 Sequential Samplers with State

Let $\mathcal{X}$ denote the output alphabet of a block, so that $|\mathcal{X}| = 2^\ell$, and let $\mathcal{S} = \{0,1\}^k$ denote the state space.

Definition C.1 (Sequential sampler with k bits of state)**.** *A sequential sampler with k bits of state and size B for n blocks over alphabet $\mathcal{X}$ is a sequence $C = (C_1, \ldots, C_n)$ of randomized circuits, each of size at most B, with state space $\mathcal{S} = \{0,1\}^k$. The variables $(X_1, S_1), \ldots, (X_n, S_n)$ are generated using independent random coins for each circuit:*

$$(X_1, S_1) \leftarrow C_1(1^n),$$

and for $i \ge 2$,

$$(X_i, S_i) \leftarrow C_i(1^n, X_{<i}, S_{i-1}).$$

The sampler C outputs $X = (X_1, \ldots, X_n)$; the states $S_1, \ldots, S_n$ are not part of its output.

To simulate C_i for $i \ge 2$ on a prefix $x_{<i}$ such that $\Pr[X_{<i} = x_{<i}] > 0$, it suffices to sample S_{i-1} conditioned on $X_{<i} = x_{<i}$, and then run $C_i(1^n, x_{<i}, S_{i-1})$. The following lemma justifies using only this state distribution, without reconstructing the preceding private states.

Lemma C.2 (Markov Chain of Sequential Sampling States)**.** *Let $(X_1, S_1), \ldots, (X_n, S_n)$ be generated by a sequential sampler with private state. For $i \in [n]$, define $A_i := (X_{\le i}, S_i)$. Then $A_1 - A_2 - \cdots - A_n$ is a Markov chain.*

Proof. Fix $i \in \{2, \ldots, n\}$ and $a_1, \ldots, a_{i-1}$ such that $\Pr[A_1 = a_1, \ldots, A_{i-1} = a_{i-1}] > 0$. Write $a_{i-1} = (x_{<i}, s_{i-1})$. For every $x_i \in \mathcal{X}$ and $s_i \in \mathcal{S}$, put $a_i = (x_{\le i}, s_i)$. Since C_i uses fresh random coins,

$$\begin{aligned}&\Pr[A_i = a_i \mid A_1 = a_1, \ldots, A_{i-1} = a_{i-1}]\\ &\quad = \Pr[C_i(1^n, x_{<i}, s_{i-1}) = (x_i, s_i)]\\ &\quad = \Pr[A_i = a_i \mid A_{i-1} = a_{i-1}].\end{aligned}$$

If the prefix in a_i does not extend $x_{<i}$, both conditional probabilities are zero. This proves the Markov property. □

C.2 Sampling from an Empirical Distribution

Before constructing a stateless sampler from a stateful one, we give a sampling circuit that reduces the circuit-size overhead. In the construction below, we run C_i M times to estimate the distribution of states supplied to C_{i+1}. We use the Laplace estimate (Definition A.10) to give every state positive probability and keep the KL error finite. Once these observations have been obtained, a direct implementation on a domain $[K] = \{1, \ldots, K\}$ uses $\widetilde{O}(M)$ gates per sample, where the tilde hides polylogarithmic factors in M and K. Producing M independent samples separately therefore costs $\widetilde{O}(M^2)$ gates. By batching these draws using a sorting network, we reduce this cost to $\widetilde{O}(M)$. The following lemma gives the circuit and bounds the additional KL error.

Lemma C.3 (Agnostic Approximate Sampling in Size Linear in Sample Count)**.** *Let $M, K \ge 1$ be integers. There exists a randomized Boolean circuit* Sample *of size $\widetilde{O}_{M,K}(M)$, where the tilde hides polylogarithmic factors in M and K, such that for any distributions D, P over $[K] = \{1, \ldots, K\}$ and independent observations $x_1, \ldots, x_M \sim D$, the circuit* Sample$(x_1, \ldots, x_M)$ *outputs M samples that, conditional on the observations, are independent with a common distribution $\widetilde{D}$ satisfying*

$$\mathop{\mathbb{E}}_{x_1,\ldots,x_M \sim D} D_{\mathsf{KL}}\left(P \parallel \widetilde{D}\right) \le D_{\mathsf{KL}}\left(P \parallel D\right) + \frac{2K}{M}.$$

Proof. We construct Sample to also accept fewer than M observations, as needed below. Fix the input observations, write $H \le M$ for their number, and let $\widehat{D}$ be their Laplace estimate, taking it to be uniform when $H = 0$. Sampling uniformly from these observations together with one copy of every element of $[K]$ gives $\widehat{D}$. We approximate this sampling using a fixed number of fair random bits and implement all M draws together by sorting.

Index the K additional elements by $0, \ldots, K-1$, followed by the H observations in their original order. Put $r = \lceil \log(2M(M+K)) \rceil$. For each $q \in [M]$, draw an independent r-bit integer U_q uniformly and set

$$J_q = \left\lfloor \frac{U_q(H+K)}{2^r} \right\rfloor.$$

The qth output of Sample is $J_q + 1$ if $J_q < K$, and the $(J_q - K + 1)$st input observation otherwise. Thus the outputs are independent conditional on the input observations and have the same distribution $\widetilde{D}$.

Each index $v \in \{0, \ldots, H+K-1\}$ has probability

$$\Pr[J_q = v] \ge \frac{\lfloor 2^r/(H+K) \rfloor}{2^r} \ge \frac{1}{H+K} - \frac{1}{2^r} \ge \frac{1 - 1/(2M)}{H+K}.$$

Summing over the indices occupied by each $z \in [K]$ gives

$$\widetilde{D}(z) \ge \left(1 - \frac{1}{2M}\right) \widehat{D}(z) > 0. \tag{4}$$

Since $(1 - 1/(2M))^M \geq 1/2$,

$$\log \frac{1}{1 - 1/(2M)} \leq \frac{1}{M}. \tag{5}$$

When the input consists of independent observations $x_1, \ldots, x_M \sim D$, Claim A.11 therefore gives

$$\begin{aligned}
\mathop{\mathbb{E}}_{x_1,\ldots,x_M \sim D} D_{\mathsf{KL}}\left(P \parallel \widetilde{D}\right) &\leq \mathop{\mathbb{E}}_{x_1,\ldots,x_M \sim D} D_{\mathsf{KL}}\left(P \parallel \widehat{D}\right) + \frac{1}{M} \\
&\leq D_{\mathsf{KL}}\left(P \parallel D\right) + \log \frac{M+K}{M+1} + \frac{1}{M} \\
&\leq D_{\mathsf{KL}}\left(P \parallel D\right) + \frac{(K-1)\log e}{M+1} + \frac{1}{M} \leq D_{\mathsf{KL}}\left(P \parallel D\right) + \frac{2K}{M}.
\end{aligned}$$

It remains to implement the M draws within the claimed circuit size. Represent the input observations by M values $z_1, \ldots, z_M \in [K]$ and bits $b_1, \ldots, b_M$: retain z_t precisely when $b_t = 1$, in increasing order of t. Encode each value $z \in [K]$ by the $\lceil \log K \rceil$-bit representation of $z - 1$. Compute $h_t = \sum_{v=1}^{t} b_v$ with $O(M \log(M+1))$ gates by successively adding the bits to a counter. Each retained occurrence has a distinct rank $h_t \in [H]$, and $H = h_M$. Computing each J_q costs $O(r^2)$ gates for multiplication, for a total of $O(Mr^2)$ gates. If $J_q < K$, its binary representation directly encodes the output $J_q + 1$. Otherwise, retrieve the occurrence of rank $J_q - K + 1$ using two sorts and a scan:

1. Form M data records containing z_t, with rank h_t if $b_t = 1$ and rank $M + 1$ otherwise. Form M query records, each carrying its original index q, including when $J_q < K$. Assign rank $J_q - K + 1$ when $J_q \geq K$ and rank $M + 1$ otherwise. Pad to a power of two with data records of rank $M + 1$ and value 1. Sort by rank, placing a data record before queries of the same rank. Every requested rank in $[H]$ is followed by its queries before any data record of larger rank.

2. Scan the sorted records, starting with value $1 \in [K]$. At a data record, replace the stored value by that record's value; at a query record, copy the stored value into the query. Thus every query with $J_q \geq K$ receives the requested observation, including when several queries have the same rank.

3. Sort by the original query index q, assigning all data records, including padding, the index $M + 1$. The first M records are the queries in their original order. Output the retrieved value when $J_q \geq K$, and $J_q + 1$ otherwise.

Both sorts operate on $O(M)$ records. Batcher's sorting networks use $O(M \log^2(M+1))$ comparators (Batcher, 1968). Each record contains a rank, a value in $[K]$, an index, and a bit distinguishing data from queries, so it has $O(r)$ bits. Each comparison and conditional swap uses $O(r)$ gates, and the scan uses $O(Mr)$ gates. Hence

$$\begin{aligned}
|\mathsf{Sample}| &= O(M \log(M+1)) + O(Mr^2) + O(Mr \log^2(M+1)) \\
&= O(M \log^3(2MK)),
\end{aligned}$$

using $r = O(\log(2MK))$. The circuit uses exactly $Mr = O(M \log(2MK))$ independent fair random bits. For every fixed input, the lookup computation implements the specified function of each U_q, as required. □

Theorem C.4 (State elimination for sequential samplers (small KL)). *Let $n \geq 1$ and $k, \ell \geq 0$ be integers, and let $0 < \delta \leq 1$. Let $C = (C_1, \ldots, C_n)$ be a sequential sampler with k bits of private state and size at most B per circuit, over alphabet $\mathcal{X} = \{0,1\}^\ell$, and put $\mathcal{S} = \{0,1\}^k$. There is a randomized stateless sequential sampler $D = (D_1, \ldots, D_n)$ such that*

$$D_{\mathsf{KL}}\left(C(1^n) \parallel D(1^n)\right) \leq \delta$$

and every D_i has Boolean circuit size at most

$$O\Big(\frac{n^3|\mathcal{X}||\mathcal{S}|}{\delta}\left[B + \log^3\Big(\frac{n|\mathcal{X}||\mathcal{S}|}{\delta}\Big)\right]\Big).$$

The circuit construction is deterministic; all sampling and estimation randomness is internal to the circuits.

The dependence on the number of states is linear up to logarithmic factors. At step j, the simulation estimates the joint distribution of (X_j, S_j) conditioned on $X_{<j} = x_{<j}$, using a fixed number of samples. Conditioning this estimate on its first component being x_j gives an estimate of the distribution of S_j conditioned on $X_{\le j} = x_{\le j}$. The KL chain rule bounds the expected KL error of this state estimate, for $x_j \sim X_j \mid X_{<j} = x_{<j}$, by the KL error of the joint estimate.

Proof. Write $X = C(1^n)$, $a = |\mathcal{X}|$, and $m = |\mathcal{S}|$. Let D_1 run $C_1(1^n)$ and return the first component of its output. Thus $D_1(1^n) \equiv X_1$, proving the case $n = 1$. Assume $n \ge 2$ and set

$$M = \left\lceil \frac{4n^2 am}{\delta} \right\rceil .$$

We use the circuit Sample from Lemma C.3 with parameters M, m, identifying $[m]$ with the state space $\mathcal{S}$. For $i \ge 2$, the circuit $D_i(1^n, x_{<i})$ runs the following algorithm, using fresh independent random bits in every call:

1. Draw M pairs from $C_1(1^n)$. Form a list L_1 of states, consisting of the second components of those symbol-state pairs whose first component equals x_1.
2. For $j = 2, \dots, i-1$, run $\mathsf{Sample}(L_{j-1})$ to obtain M states $s_1, \dots, s_M$. For each $t \in [M]$, run $C_j(1^n, x_{<j}, s_t)$. Form a list L_j from the second components of those symbol-state pairs whose first component equals x_j.
3. Run $\mathsf{Sample}(L_{i-1})$, take its first output s, and return the first component of $C_i(1^n, x_{<i}, s)$.

Each list retains repeated states in their original order and may be empty. As in the lemma's construction, it is represented by M states and M bits specifying which observations are retained. All lists are computed inside D_i; no state is passed between the circuits $D_1, \dots, D_n$.

In the following, $[\cdot]_1$ denotes the first output component, and, for $\Pr[X_{\le j} = x_{\le j}] > 0$,

$$\begin{aligned}
\pi_j^{x_{\le j}}(s) &= \Pr[S_j = s \mid X_{\le j} = x_{\le j}], \\
\widetilde{\pi}_j^{x_{\le j}}(s) &= \Pr\big[[\mathsf{Sample}(L_j)]_1 = s \mid L_j\big], \\
\varepsilon_j &= \mathop{\mathbb{E}}_{x_{\le j} \sim X_{\le j}} \mathop{\mathbb{E}}_{L_j \mid x_{\le j}} D_{\mathsf{KL}}\left(\pi_j^{x_{\le j}} \,\|\, \widetilde{\pi}_j^{x_{\le j}}\right),
\end{aligned}$$

and the expectation $\mathbb{E}_{L_j \mid x_{\le j}}$ averages over the algorithm's random bits on input $x_{\le j}$. We have,

$$\begin{aligned}
&D_{\mathsf{KL}}\left(C(1^n) \,\|\, D(1^n)\right) \\
&= \sum_{i=2}^{n} \mathop{\mathbb{E}}_{x_{<i} \sim X_{<i}} D_{\mathsf{KL}}\left(X_i \mid_{X_{<i} = x_{<i}} \,\|\, D_i(1^n, x_{<i})\right) \\
&= \sum_{i=2}^{n} \mathop{\mathbb{E}}_{x_{<i} \sim X_{<i}} D_{\mathsf{KL}}\left(X_i \mid_{X_{<i} = x_{<i}} \,\|\, \left[C_i(1^n, x_{<i}, [\mathsf{Sample}(L_{i-1})]_1)\right]_1\right) \\
&\le \sum_{i=2}^{n} \mathop{\mathbb{E}}_{x_{<i} \sim X_{<i}} \mathop{\mathbb{E}}_{L_{i-1} \mid x_{<i}} D_{\mathsf{KL}}\left(\pi_{i-1}^{x_{<i}} \,\|\, \widetilde{\pi}_{i-1}^{x_{<i}}\right) = \sum_{j=1}^{n-1} \varepsilon_j .
\end{aligned}$$

The equalities use the KL chain rule and step 3; the inequality uses Lemma C.2, data processing, and convexity, since D_i constructs L_{i-1} using fresh randomness.

For $2 \le j \le n-1$, fix $x_{<j}$ with $\Pr[X_{<j} = x_{<j}] > 0$ and the computed list L_{j-1}. Set

$$\begin{aligned}
P_j(x, s) &= \Pr[X_j = x, S_j = s \mid X_{<j} = x_{<j}], \\
Q_j(x, s) &= \Pr\big[C_j(1^n, x_{<j}, [\mathsf{Sample}(L_{j-1})]_1) = (x, s)\big].
\end{aligned}$$

Step 2 produces M independent pairs $(Z_t, T_t)_{t=1}^{M}$ from Q_j.

$$L_j = (T_t : Z_t = x_j),$$
$$N_{x,s} = \sum_{t=1}^{M} \mathbf{1}[(Z_t, T_t) = (x, s)], \qquad N_x = \sum_{s \in \mathcal{S}} N_{x,s},$$
$$\widehat{Q}_j(x, s) = \frac{N_{x,s} + 1}{M + am}, \qquad \widehat{Q}_j(s \mid x_j) = \frac{N_{x_j,s} + 1}{N_{x_j} + m}.$$

By Equation (4),

$$\widetilde{\pi}_j^{x_{\le j}}(s) \ge \left(1 - \frac{1}{2M}\right) \widehat{Q}_j(s \mid x_j) > 0. \tag{6}$$

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{(Z_t,T_t)\sim Q_j \\ t\in[M]}} \mathop{\mathbb{E}}_{x_j \sim X_j \mid X_{<j}=x_{<j}} D_{\mathsf{KL}}\left(\pi_j^{x_{\le j}} \,\|\, \widetilde{\pi}_j^{x_{\le j}}\right) \\
&\le \mathop{\mathbb{E}}_{\substack{(Z_t,T_t)\sim Q_j \\ t\in[M]}} \mathop{\mathbb{E}}_{x_j \sim X_j \mid X_{<j}=x_{<j}} D_{\mathsf{KL}}\left(\pi_j^{x_{\le j}} \,\|\, \widehat{Q}_j(\cdot \mid x_j)\right) + \frac{1}{M} \\
&\le \mathop{\mathbb{E}}_{\substack{(Z_t,T_t)\sim Q_j \\ t\in[M]}} D_{\mathsf{KL}}\left(P_j \,\|\, \widehat{Q}_j\right) + \frac{1}{M} \\
&\le D_{\mathsf{KL}}\left(P_j \,\|\, Q_j\right) + \log\frac{M + am}{M + 1} + \frac{1}{M} \\
&\le D_{\mathsf{KL}}\left(\pi_{j-1}^{x_{<j}} \,\|\, \widetilde{\pi}_{j-1}^{x_{<j}}\right) + \frac{2am}{M}.
\end{aligned}$$

The successive bounds are Equation (5), the KL chain rule, Claim A.11, and data processing through C_j, with $\log((M + am)/(M + 1)) + 1/M \le 2am/M$. For $j = 1$, step 1 gives $P_1 = Q_1 = C_1(1^n)$, so the same calculation ends at $2am/M$. For $j \ge 2$, averaging over $x_{<j} \sim X_{<j}$ and the list L_{j-1} computed on that prefix gives

$$\begin{aligned}
&\varepsilon_1 \le \frac{2am}{M}, \\
&\varepsilon_j \le \varepsilon_{j-1} + \frac{2am}{M} && (2 \le j \le n-1), \\
&\varepsilon_j \le \frac{2jam}{M} && (1 \le j \le n-1).
\end{aligned}$$

$$\begin{aligned}
D_{\mathsf{KL}}\left(C(1^n) \,\|\, D(1^n)\right) &\le \sum_{j=1}^{n-1} \varepsilon_j \le \frac{2am}{M} \sum_{j=1}^{n-1} j \\
&= \frac{n(n-1)am}{M} \le \delta.
\end{aligned} \tag{7}$$

Boolean circuit size. Each call to Sample costs $O(M \log^3(2Mm))$ gates by Lemma C.3. Forming a list uses M copies of C_j, each of size at most B, and M comparisons with x_j, each costing $O(\ell)$ gates. The returned states and comparison bits are exactly the list representation used by Sample. There are $i - 1$ such lists, $i - 1$ calls to Sample, and one final call to C_i. Thus

$$\begin{aligned}
|D_i| &= O\big(iM\left[B + \ell + \log^3(2Mm)\right]\big) \\
&= O\left(\frac{n^3 am}{\delta}\left[B + \log^3\left(\frac{2nam}{\delta}\right)\right]\right),
\end{aligned}$$

since $i \le n$, $M = O(n^2 am/\delta)$, and $\ell = \log a = O(\log(2Mm))$. All calls and list representations have fixed size, so this construction gives a stateless randomized Boolean circuit. □

C.3 Sample Complexity of Estimating the State Distribution

The construction in Theorem C.4 estimates the entire conditional state distribution before sampling a state from that estimate. We show that, when this distribution is unknown and the only access to it is through independent samples, expected KL divergence at most $1/4$ requires $\Omega(|\mathcal{S}|) = \Omega(2^k)$ samples. This establishes the necessity of the linear dependence on $|\mathcal{S}|$ when estimating the state distribution from independent samples.

Proposition C.5 (Sample Lower Bound for Estimating the State Distribution)**.** *Let $k \geq 1$ and $N \geq 0$ be integers, and let $\mathcal{S} = \{0,1\}^k$. Suppose that a randomized algorithm receives at most N independent samples from an unknown distribution P over $\mathcal{S}$ and outputs a vector of probabilities $\hat{\pi}$ over $\mathcal{S}$, with no subsequent access to the sampling oracle. The algorithm receives no other information about P. Then there is a distribution P for which*

$$\mathbb{E}\big[D_{\mathsf{KL}}\left(P \parallel \hat{\pi}\right)\big] \geq 1 - \frac{2N}{|\mathcal{S}|},$$

where the expectation is over the samples and the algorithm's internal randomness. Consequently, if $\mathbb{E}[D_{\mathsf{KL}}\left(P \parallel \hat{\pi}\right)] \leq 1/4$ for every P, then $N \geq 3|\mathcal{S}|/8 = 3 \cdot 2^{k-3}$.

Proof. We partition the states into pairs and choose one state from each pair using a hidden bit. Each sample reveals only the bit for its pair; every pair that is never sampled contributes to the expected KL error.

Let $m = 2^{k-1}$ and identify $\mathcal{S}$ with $[m] \times \{0,1\}$. For each $\theta = (\theta_1, \ldots, \theta_m) \in \{0,1\}^m$, define

$$P_\theta(a,b) = \begin{cases} 1/m, & b = \theta_a, \\ 0, & b \neq \theta_a. \end{cases}$$

Choose Θ uniformly from $\{0,1\}^m$. Conditional on Θ, a sample from P_Θ is generated by choosing A uniformly from $[m]$, independently of Θ, and returning (A, Θ_A). We may generate N independent samples in advance and let the algorithm read as many as it requests, ignoring any unused samples.

Let W consist of these N samples and the algorithm's independent random coins. Fix W, so that the output $\hat{\pi}$ is fixed, and let $U \subseteq [m]$ contain the first coordinates absent from the samples. For each $a \notin U$, the samples determine Θ_a; conditional on W, the bits $(\Theta_a)_{a \in U}$ remain independent and uniform. Consequently, the distribution $\bar{P} = \mathbb{E}[P_\Theta \mid W]$ satisfies

$$\bar{P}(a,b) = \begin{cases} 1/m, & a \notin U \text{ and } b = \Theta_a, \\ 0, & a \notin U \text{ and } b \neq \Theta_a, \\ 1/(2m), & a \in U. \end{cases}$$

In particular,

$$H(\bar{P}) = \frac{m - |U|}{m} \log m + \frac{|U|}{m} \log(2m) = \log m + \frac{|U|}{m}, \qquad H(P_\theta) = \log m.$$

Expanding KL divergence gives

$$\begin{aligned} \mathbb{E}\big[D_{\mathsf{KL}}\left(P_\Theta \parallel \hat{\pi}\right) \mid W\big] &= -\log m - \sum_{s \in \mathcal{S}} \bar{P}(s) \log \hat{\pi}(s) \\ &= D_{\mathsf{KL}}\left(\bar{P} \parallel \hat{\pi}\right) + H(\bar{P}) - \log m \geq \frac{|U|}{m}. \end{aligned}$$

At most N pairs appear among the samples, so $|U| \geq m - N$. Therefore

$$\mathbb{E}\big[D_{\mathsf{KL}}\left(P_\Theta \parallel \hat{\pi}\right)\big] \geq \frac{\mathbb{E}\,|U|}{m} \geq 1 - \frac{N}{m} = 1 - \frac{2N}{|\mathcal{S}|}.$$

Averaging over the uniform choice of Θ implies that some fixed θ satisfies the same lower bound, with expectation only over its samples and the algorithm's coins. If this expectation is at most $1/4$ for every P_θ, then $1 - 2N/|\mathcal{S}| \leq 1/4$, or $N \geq 3|\mathcal{S}|/8$. □

Every distribution P_θ used in the proof can occur as a conditional state distribution: the first circuit can output a fixed block x_1, choose A uniformly from $[2^{k-1}]$, and set $S_1 = (A, \theta_A)$. Then $\pi_1^{x_1} = P_\theta$. Thus the lower bound applies even when independent samples from $\pi_1^{x_1} = P_\theta$ are supplied directly. The first circuit may encode θ in its description; the proposition assumes access only to samples, rather than to that description.

The lower bound concerns estimating a probability distribution from samples. It does not establish a matching circuit-size lower bound for Theorem C.4, or a lower bound for state elimination procedures that do not estimate the entire state distribution. In particular, obtaining one state with distribution P from an oracle that samples P takes one query; if $\hat{\pi}$ is the point mass at that sample, then $\mathbb{E}[\hat{\pi}] = P$, but $\mathbb{E}[D_{\mathsf{KL}}(P \parallel \hat{\pi})] = +\infty$ whenever P has at least two support points.

D Insertional Sampling and Insertional Pseudoentropy

Definition D.1 (Stateless Insertional Sampler)**.** *Let $n, \ell \geq 1$ and let L be a distribution over $(\{0,1\}^\ell)^n$. An insertional sampler $C = (C_1, \ldots, C_n)$ of L is a sequence of randomized (possibly unbounded) circuits, where*

- *$C(1^n) \coloneqq C_n(C_{n-1}(\cdots(C_1(\emptyset))\cdots)) \equiv L$, and*
- *for every $i \in [n]$ and $x \in (\{0,1\}^\ell)^{i-1}$, $C_i(x)$ is a distribution over strings in $(\{0,1\}^\ell)^i$ for which x is a subsequence.*

We say that C has size B if for all $i \in [n]$, C_i is of size at most B.

Definition D.2 (Insertional decomposition)**.** *Let $n, \ell \geq 1$ and let L be a distribution over $(\{0,1\}^\ell)^n$. An insertional decomposition $(L_0, \ldots, L_n)$ of L is a sequence of jointly distributed random variables, where*

- *L_0 is the fixed distribution over the empty string $\emptyset$,*
- *$L_n \equiv L$, and*
- *for every $i \in [n]$ and $x \in (\{0,1\}^\ell)^{i-1}$ with $\Pr[L_{i-1} = x] > 0$, $L_i \mid L_{i-1} = x$ is a distribution over strings in $(\{0,1\}^\ell)^i$ for which x is a subsequence.*

For each insertional decomposition, we compare the apparent conditional entropy of each insertion with its actual conditional entropy. The randomness in choosing the insertion order therefore does not itself contribute to the pseudoentropy gap.

Definition D.3 (Insertional Pseudoentropy)**.** *Let $n, \ell \geq 1$ and let L be distributed over $(\{0,1\}^\ell)^n$. Let $T \geq 1$, $k \geq 0$, and $0 < \varepsilon < 1$; these quantities may depend on a security parameter λ. We say that L has (T, ε) insertional pseudoentropy at least k if, for every insertional decomposition $(L_0, \ldots, L_n)$ of L, with I uniform on $[n]$ independently of the decomposition, there exists a random variable Z, jointly distributed with $(I, L_0, \ldots, L_n)$ and satisfying $Z \in (\{0,1\}^\ell)^I$ whenever $I = i$, such that*

1. *$H(Z \mid I, L_{I-1}) - H(L_I \mid I, L_{I-1}) \geq \frac{k - H(L)}{n}$, and*
2. *(I, L_I, L_{I-1}) and (I, Z, L_{I-1}) are (T, ε) nonuniformly indistinguishable.*

The insertional pseudoentropy is the supremum of the values k satisfying this definition.

Remark D.4 (Comparison with next-block pseudoentropy)**.** Both definitions measure the increase in conditional entropy that can remain computationally undetectable. In next-block pseudoentropy, the entropy chain rule makes the subtraction of the actual conditional entropy implicit. For $Y = (Y_1, \ldots, Y_n)$ with $Y_i \in \{0,1\}^\ell$, $\ell \geq 1$, and I uniform on $[n]$ independently of Y, the length of $Y_{<I}$ determines I, so

$$nH(Y_I \mid Y_{<I}) = \sum_{i=1}^{n} H(Y_i \mid Y_{<i}) = H(Y).$$

Consequently, the condition $H(Z \mid Y_{<I}) \geq k/n$ in Definition A.15 is equivalent to

$$H(Z \mid Y_{<I}) - H(Y_I \mid Y_{<I}) \geq \frac{k - H(Y)}{n}.$$

This is the same entropy-gap form as in Definition D.3. For an insertional decomposition, however, the entropy chain rule gives

$$\begin{aligned} nH(L_I \mid I, L_{I-1}) &= \sum_{i=1}^{n} H(L_i \mid L_{i-1}) \\ &= \sum_{i=1}^{n} \big(H(L_i) - H(L_{i-1}) + H(L_{i-1} \mid L_i)\big) \\ &= H(L) + \sum_{i=1}^{n} H(L_{i-1} \mid L_i). \end{aligned}$$

In next-block pseudoentropy, a prefix determines all prior information, whereas an in insertional pseudoentropy, a string given by L_i does not necessarily determine its preceding string L_{i-1}. For example, for even i, if L_i is 01 repeated $i/2$ times, then there are i distinct possibilities for L_{i-1} and a random choice of L_{i-1} among them would result in $H(L_{i-1} \mid L_i) = \log i$. Therefore, the sum $\sum_{i\in[n]} H(L_{i-1} \mid L_i)$ could be as large as $\log(n!)$. Subtracting the actual conditional entropy in Definition D.3 ensures that the gap $k - H(L)$ does not count randomness in the choice of insertion order. The quantification over every insertional decomposition accounts for the sampler's freedom to choose this order. This is implicitly the case for next-block pseudoentropy since left-to-right generation has a single fixed prefix decomposition.

The *insertional pseudoentropy gap* is the insertional pseudoentropy minus $H(L)$. Taking $Z = L_I$ shows that this gap is nonnegative. The joint distribution of Z is chosen for each decomposition before the distinguishing circuit: the same Z must fool every Boolean circuit of size at most T. The guarantee also holds for randomized circuits with independent coins, by fixing their coins and averaging. The length of L_{I-1} determines I, so conditioning on (I, L_{I-1}) is equivalent to conditioning on L_{I-1}. The variable Z need not be a valid one-symbol extension of L_{I-1}.

Remark D.5 (Equivalent formulation)**.** Equivalently, every insertional decomposition $(L_0, \ldots, L_n)$ admits random variables $Z_i \in (\{0,1\}^\ell)^i$, jointly distributed with the decomposition and independently of I, satisfying

$$\sum_{i=1}^{n} \big(H(Z_i \mid L_{i-1}) - H(L_i \mid L_{i-1})\big) \geq k - H(L)$$

and making (I, Z_I, L_{I-1}) and (I, L_I, L_{I-1}) (T, ε) nonuniformly indistinguishable. The family is chosen once for all distinguishing circuits.

Given Z, draw each Z_i conditionally on the decomposition using the distribution of Z given $I = i$ and L_{i-1}, with fresh randomness independent of I. On zero-probability conditioning events, choose any distribution on $(\{0,1\}^\ell)^i$. Conversely, given such a family, take $Z = Z_I$ with I independent of the decomposition and the family. These constructions preserve the joint distributions being compared and give

$$\begin{aligned} H(Z \mid I, L_{I-1}) &= \frac{1}{n} \sum_{i=1}^{n} H(Z_i \mid L_{i-1}), \\ H(L_I \mid I, L_{I-1}) &= \frac{1}{n} \sum_{i=1}^{n} H(L_i \mid L_{i-1}). \end{aligned}$$

Thus both formulations have the same entropy gap, circuit size, and distinguishing error.

D.1 Characterization of Insertional Samplability

Theorem D.6 (Characterization of stateless insertional sampling)**.** *There exist a universal constant $c_0 \in (0, 1]$ and a polynomial p with nonnegative coefficients and constant term at least one such that the following holds. Let $n, \ell \geq 1$ and let L take values in $(\{0,1\}^\ell)^n$. Let $t \geq 1$, $0 < \epsilon, \eta, \sigma < 1$, and $\delta \geq \epsilon$.*

1. *If L has (t, ϵ) insertional pseudoentropy at least $H(L)+\delta$, then, for every $B \geq 1$ satisfying*
$$B \leq \frac{\min\{t^{c_0},\, \sigma/(n\epsilon)\}}{p(n, 2^\ell, 1/\eta, \log(n/\sigma))},$$
every size-B stateless insertional sampler C over $\{0,1\}^\ell$ satisfies
$$D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) > \delta - \sigma - n\eta.$$

2. *If the (t, ϵ) insertional pseudoentropy of L is at most $H(L) + \delta - \epsilon$, then there exists a size-B stateless insertional sampler C over $\{0,1\}^\ell$ with*
$$B \leq p(t, n, 2^\ell, 1/\epsilon), \qquad D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) \leq \delta.$$

Remark D.7. If $\delta < \epsilon$ and some $B \geq 1$ satisfies the size bound in item 1, then $p \geq 1$ forces $\sigma \geq n\epsilon > \delta$, so the claimed lower bound is negative and the inequality follows from nonnegativity of KL-divergence. If no such B exists, item 1 is vacuous.

Lemma D.8 (Distinct one-symbol insertions)**.** *Let $i, \ell \geq 1$ and $w \in (\{0,1\}^\ell)^{i-1}$. The set*
$$\mathcal{E}_i(w) = \{z \in (\{0,1\}^\ell)^i : w \text{ is a subsequence of } z\}$$
has exactly $K_i = i(2^\ell - 1) + 1$ elements and admits a bijection $\mathsf{Ext}_i(w, \cdot) : [K_i] \to \mathcal{E}_i(w)$. Given (i, ℓ, w, k), one can compute $\mathsf{Ext}_i(w, k)$ in $\mathrm{poly}(i, 2^\ell)$ time. Given (i, ℓ, w, z) with $z \in \mathcal{E}_i(w)$, one can compute the unique $k \in [K_i]$ satisfying $\mathsf{Ext}_i(w, k) = z$ in the same time bound.

Proof. We count each resulting string once by choosing the rightmost position at which its inserted symbol can be placed. For $a \in \{0,1\}^\ell$ and $j \in [i]$, write
$$\mathsf{Ins}(w, a, j) = (w_1, \ldots, w_{j-1}, a, w_j, \ldots, w_{i-1}).$$
For $a, b \in \{0,1\}^\ell$ and $1 \leq j < k \leq i$, comparison of coordinates $j, \ldots, k$ gives
$$\mathsf{Ins}(w, a, j) = \mathsf{Ins}(w, b, k) \quad \Longleftrightarrow \quad a = w_j = \cdots = w_{k-1} = b.$$
Retain all pairs (a, j) with $j = i$, and those with $j < i$ and $a \neq w_j$. By the displayed equivalence, these are exactly the rightmost representations, one for each output. Hence
$$|\mathcal{E}_i(w)| = 2^\ell + (i-1)(2^\ell - 1) = i(2^\ell - 1) + 1.$$
Order the retained pairs by increasing j, then lexicographically by a. Define $\mathsf{Ext}_i(w, k)$ by inserting the symbol at the position specified by the k-th pair. To invert this map given w and z, enumerate the retained pairs and return the index of the unique pair whose insertion produces z. Both algorithms examine at most $i2^\ell$ pairs and use $O(i\ell)$ time per pair, giving the claimed $\mathrm{poly}(i, 2^\ell)$ bound. □

Lemma D.9 (Conditional and final KL-divergence)**.** *Let $n, \ell \geq 1$ and let L be distributed over $(\{0,1\}^\ell)^n$. For every sequence $C = (C_1, \ldots, C_n)$ of valid insertion circuits,*
$$\inf_{(L_0, \ldots, L_n)} \sum_{i=1}^{n} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) = D_{\mathsf{KL}}\left(L \parallel C(1^n)\right),$$
where the infimum is over insertional decompositions of L. Moreover, if $D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) < \infty$, there exists an insertional decomposition $(L_0, \ldots, L_n)$ of L that forms a Markov chain and satisfies
$$\sum_{i=1}^{n} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) = D_{\mathsf{KL}}\left(L \parallel C(1^n)\right).$$

Proof. Let $(L_0, \ldots, L_n)$ be any insertional decomposition of L, and $X_0 = \emptyset$, $X_i \leftarrow C_i(X_{i-1})$ with fresh independent coins, so $X_n \equiv C(1^n)$. Then,
$$\begin{aligned}
D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) &= D_{\mathsf{KL}}\left(L_n \parallel X_n\right) \\
&\leq D_{\mathsf{KL}}\left((L_{n-1}, L_n) \parallel (X_{n-1}, X_n)\right) \\
&= D_{\mathsf{KL}}\left(L_{n-1} \parallel X_{n-1}\right) + \mathop{\mathbb{E}}_{x \sim L_{n-1}} D_{\mathsf{KL}}\left(L_n \mid_{L_{n-1}=x} \parallel C_n(x)\right) \\
&\leq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{x \sim L_{i-1}} D_{\mathsf{KL}}\left(L_i \mid_{L_{i-1}=x} \parallel C_i(x)\right) \\
&= \sum_{i=1}^{n} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right).
\end{aligned}$$

The first inequality is Fact A.9 applied to the map that keeps the second coordinate. The next equality is Fact A.7: given $X_{n-1} = x$, the distribution of X_n is $C_n(x)$. Repeating these two steps down to $i = 1$ gives the sum, since $D_{\mathsf{KL}}(L_0 \parallel X_0) = 0$. The last equality applies the same chain rule to two pairs with the same first-coordinate distribution L_{i-1}. Since the decomposition was arbitrary, the final KL divergence is at most the infimum in the statement.

Write $Q(x_0, \ldots, x_n) = \Pr[X_0 = x_0, \ldots, X_n = x_n]$ and $Q_n(z) = \Pr[X_n = z]$. For the reverse inequality, assume $D_{\mathsf{KL}}(L \parallel Q_n) < \infty$ and define

$$M(x_0, \ldots, x_n) = \begin{cases} Q(x_0, \ldots, x_n) \dfrac{\Pr[L = x_n]}{Q_n(x_n)}, & Q_n(x_n) > 0, \\ 0, & Q_n(x_n) = 0. \end{cases}$$

For $Q_n(z) > 0$,

$$\begin{aligned} \sum_{x_0, \ldots, x_{n-1}} M(x_0, \ldots, x_{n-1}, z) &= \frac{\Pr[L = z]}{Q_n(z)} \sum_{x_0, \ldots, x_{n-1}} Q(x_0, \ldots, x_{n-1}, z) \\ &= \Pr[L = z]. \end{aligned}$$

Both sides are zero otherwise, since $\mathsf{supp}(L) \subseteq \mathsf{supp}(Q_n)$. Thus $(M_0, \ldots, M_n) \sim M$ satisfies $M_n \equiv L$. Since $\mathsf{supp}(M) \subseteq \mathsf{supp}(Q)$, $M_0 = \emptyset$ and every step inserts one symbol.

To compute the conditional probabilities, for $i \in \{0, \ldots, n\}$ set

$$h_i(w) = \sum_{z:\, Q_n(z) > 0} \Pr[X_n = z \mid X_i = w] \frac{\Pr[L = z]}{Q_n(z)} \qquad (w \in \mathsf{supp}(X_i)),$$

and set $h_i(w) = 0$ otherwise. The Markov property of $(X_0, \ldots, X_n)$ gives

$$\begin{aligned} &\Pr[M_0 = x_0, \ldots, M_i = x_i] \\ &\quad = \sum_{z:\, Q_n(z) > 0} \Pr[X_0 = x_0, \ldots, X_i = x_i, X_n = z] \frac{\Pr[L = z]}{Q_n(z)} \\ &\quad = \Pr[X_0 = x_0, \ldots, X_i = x_i] h_i(x_i). \end{aligned}$$

For $i \in [n]$ and every tuple with $\Pr[M_0 = x_0, \ldots, M_{i-1} = x_{i-1}] > 0$, division gives

$$\begin{aligned} &\Pr[M_i = z \mid M_0 = x_0, \ldots, M_{i-1} = x_{i-1}] \\ &\quad = \Pr[C_i(x_{i-1}) = z] \frac{h_i(z)}{h_{i-1}(x_{i-1})} \\ &\quad = \Pr[M_i = z \mid M_{i-1} = x_{i-1}]. \end{aligned}$$

The ratio depends on the preceding tuple only through x_{i-1}, so the last equality holds and $(M_0, \ldots, M_n)$ is a Markov insertional decomposition of L. By Fact A.8,

$$\begin{aligned} &\sum_{i=1}^{n} D_{\mathsf{KL}}\left((M_{i-1}, M_i) \parallel (M_{i-1}, C_i(M_{i-1}))\right) \\ &\quad = D_{\mathsf{KL}}(M \parallel Q) \\ &\quad = \sum_{(x_0, \ldots, x_n) \in \mathsf{supp}(M)} M(x_0, \ldots, x_n) \log \frac{\Pr[L = x_n]}{Q_n(x_n)} \\ &\quad = \sum_{z \in \mathsf{supp}(L)} \Pr[L = z] \log \frac{\Pr[L = z]}{Q_n(z)} \\ &\quad = D_{\mathsf{KL}}(L \parallel C(1^n)). \end{aligned}$$

This decomposition attains the infimum. □

Lemma D.10 (Insertional pseudoentropy implies conditional sampling hardness)**.** *Under the hypotheses and size condition of the first part of Theorem D.6, every insertional decomposition $(L_0, \ldots, L_n)$ of L and every size-B stateless insertional sampler C over $\{0,1\}^\ell$ satisfy*

$$\mathop{\mathbb{E}}_{i \leftarrow [n]} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) > \frac{\delta - \sigma}{n} - \eta.$$

Proof. Fix an insertional decomposition $(L_0, \dots, L_n)$ and a sampler C, and put

$$d_i = D_{\mathsf{KL}}\left((L_{i-1}, L_i) \,\|\, (L_{i-1}, C_i(L_{i-1}))\right), \qquad d = \sum_{i=1}^{n} d_i.$$

If $d = \infty$, the claim follows, so assume $d < \infty$. By Definition D.3 and Remark D.5, choose a family $Z_i \in (\{0,1\}^\ell)^i$ such that

$$\sum_{i=1}^{n} H(Z_i \mid L_{i-1}) \geq \sum_{i=1}^{n} H(L_i \mid L_{i-1}) + \delta$$

and (I, Z_I, L_{I-1}) and (I, L_I, L_{I-1}) are (t, ϵ) nonuniformly indistinguishable, where I is uniform on $[n]$ and independent of the family and decomposition. We estimate the sampler's output probabilities and test their negative logarithms. The expected negative logarithm will be at least $H(Z_i \mid L_{i-1})$ on Z_i and at most $H(L_i \mid L_{i-1}) + d_i + \eta/4$ on L_i. Indistinguishability compares their averages over the uniform index I.

Let $\mathcal{E}_i(w)$ and $K_i = i(2^\ell - 1) + 1$ be as in Lemma D.8, and set

$$q = n2^\ell, \qquad M = \lceil 32q/\eta \rceil, \qquad \theta = 2^{-\lceil \log(32/\eta) \rceil}, \qquad \Lambda = n\ell + \log(1/\theta).$$

On input (i, z, w), draw M samples from $C_i(w)$ with independent coins, also independent of the decomposition and auxiliary family. Let N_z count the samples equal to z. Define

$$\widehat{P}_w(z) = \begin{cases} \dfrac{N_z + 1}{M + K_i}, & z \in \mathcal{E}_i(w), \\ 0, & z \notin \mathcal{E}_i(w), \end{cases} \qquad R_w(z) = (1-\theta)\widehat{P}_w(z) + \frac{\theta}{2^{i\ell}}.$$

Since $\sum_{z \in \mathcal{E}_i(w)} N_z = M$, both $\widehat{P}_w$ and R_w are probability distributions. The uniform term is needed because Z_i need not belong to $\mathcal{E}_i(L_{i-1})$. For every length-i string z,

$$\frac{\theta}{2^{i\ell}} \leq R_w(z) \leq 1, \qquad 0 \leq -\log R_w(z) \leq \Lambda.$$

Set $b = \lceil \log(16\Lambda/\eta) \rceil$ and $\rho = 2^{-b}$. The circuit D accepts with probability

$$\rho \left\lfloor \frac{-\log R_w(z)}{\Lambda\rho} \right\rfloor.$$

For each i, store the displayed floor for the $M+1$ valid-extension counts $N_z = 0, \dots, M$ and for the single invalid-extension case. This gives $n(M+2)$ entries, each an integer in $\{0, \dots, 2^b\}$ stored using $b+1$ bits. Draw a fresh uniform b-bit integer and accept exactly when it is smaller than the stored integer. The circuit contains M copies of each C_j; it selects the count from the branch $j = i$, whose copies receive w; the other branches receive fixed strings of the required lengths. Membership in $\mathcal{E}_i(w)$ is tested by deleting each position of z and comparing the resulting string with w. Selection, counting, these comparisons, and table lookup require $\mathrm{poly}(n, M, \ell, b)$ additional gates. Consequently, for a universal polynomial A,

$$|D| \leq nMB + \mathrm{poly}(n, M, \ell, b) \leq BA(n, 2^\ell, 1/\eta),$$

where the last inequality uses $B \geq 1$ and the choices of M, b. Choose the polynomial p in Theorem D.6 large enough that $p(n, 2^\ell, 1/\eta, \log(n/\sigma)) \geq \max\{A(n, 2^\ell, 1/\eta), 2\Lambda\}$. The assumed size bound and $c_0 \leq 1$ give

$$|D| \leq BA(n, 2^\ell, 1/\eta) \leq t^{c_0} \leq t, \qquad n\Lambda\epsilon \leq \sigma/2, \qquad \Lambda\rho \leq \eta/16.$$

We first bound the expected negative logarithm on L_i. For $w \in \mathsf{supp}(L_{i-1})$, apply Claim A.11 on $\mathcal{E}_i(w)$ with $P = C_i(w)$, $Q = L_i \mid_{L_{i-1}=w}$, and M samples. Since $R_w(z) \geq (1-\theta)\widehat{P}_w(z)$,

$$\begin{aligned} \mathbb{E}\, D_{\mathsf{KL}}\left(L_i \mid_{L_{i-1}=w} \,\|\, R_w\right) &\leq \mathbb{E}\, D_{\mathsf{KL}}\left(L_i \mid_{L_{i-1}=w} \,\|\, \widehat{P}_w\right) - \log(1-\theta) \\ &\leq D_{\mathsf{KL}}\left(L_i \mid_{L_{i-1}=w} \,\|\, C_i(w)\right) + \log\frac{M+K_i}{M+1} - \log(1-\theta). \end{aligned}$$

Here expectations are over the M samples. Since $K_i \le q$ and $\theta \le \eta/32$, the parameter choices give

$$\log \frac{M+K_i}{M+1} - \log(1-\theta) \le \left(\frac{\eta}{32} + \frac{\eta}{31}\right) \log e < \eta/4.$$

Drawing the samples independently of L_i conditional on L_{i-1} and applying Fact A.7 to d_i gives

$$\begin{aligned}\mathbb{E}[-\log R_{L_{i-1}}(L_i)] &= H(L_i \mid L_{i-1}) + \mathop{\mathbb{E}}_{w \sim L_{i-1}} \mathbb{E}\, D_{\mathsf{KL}}\left(L_i \mid_{L_{i-1}=w} \parallel R_w\right) \\ &\le H(L_i \mid L_{i-1}) + d_i + \eta/4.\end{aligned}$$

The same samples are independent of Z_i conditional on L_{i-1}. For every $w \in \mathsf{supp}(L_{i-1})$ and every value of R_w,

$$\begin{aligned}\mathop{\mathbb{E}}_{z \sim Z_i|_{L_{i-1}=w}} [-\log R_w(z)] &= H(Z_i \mid L_{i-1} = w) + D_{\mathsf{KL}}\left(Z_i \mid_{L_{i-1}=w} \parallel R_w\right) \\ &\ge H(Z_i \mid L_{i-1} = w).\end{aligned}$$

For $V_i = L_i$ or $V_i = Z_i$, the rounding gives

$$\left|\Lambda \Pr[D(i, V_i, L_{i-1}) = 1] - \mathbb{E}[-\log R_{L_{i-1}}(V_i)]\right| \le \Lambda\rho.$$

Here the expectations include the fresh samples drawn by D. Since $|D| \le t$, indistinguishability gives

$$\left|\sum_i \Pr[D(i, Z_i, L_{i-1}) = 1] - \sum_i \Pr[D(i, L_i, L_{i-1}) = 1]\right| \le n\epsilon.$$

Combining these inequalities,

$$\begin{aligned}\sum_i H(L_i \mid L_{i-1}) + \delta &\le \sum_i H(Z_i \mid L_{i-1}) \\ &\le \sum_i \mathbb{E}[-\log R_{L_{i-1}}(Z_i)] \\ &\le \Lambda \sum_i \Pr[D(i, Z_i, L_{i-1}) = 1] + n\Lambda\rho \\ &\le \Lambda \sum_i \Pr[D(i, L_i, L_{i-1}) = 1] + n\Lambda\epsilon + n\Lambda\rho \\ &\le \sum_i \mathbb{E}[-\log R_{L_{i-1}}(L_i)] + n\Lambda\epsilon + 2n\Lambda\rho \\ &\le \sum_i H(L_i \mid L_{i-1}) + d + n\eta/4 + n\Lambda\epsilon + 2n\Lambda\rho \\ &\le \sum_i H(L_i \mid L_{i-1}) + d + \sigma/2 + 3n\eta/8.\end{aligned}$$

Subtracting $\sum_i H(L_i \mid L_{i-1})$ from both sides gives

$$d \ge \delta - \sigma/2 - 3n\eta/8 > \delta - \sigma - n\eta.$$

Dividing by n proves the claim. □

Lemma D.11 (Small insertional pseudoentropy gives conditional samplers)**.** *Under the hypotheses of the second part of Theorem D.6, there are an insertional decomposition $(L_0, \ldots, L_n)$ of L and a size-B stateless insertional sampler C over $\{0,1\}^\ell$, with $B \le p(t, n, 2^\ell, 1/\epsilon)$, such that*

$$\mathop{\mathbb{E}}_{i \leftarrow [n]} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) \le \frac{\delta}{n}.$$

Proof. By assumption, the (t, ϵ) insertional pseudoentropy of L is at most $H(L) + \delta - \epsilon$. Indexing the valid extensions in Lemma D.8 gives alphabets of size at most $n2^\ell$ for applying item 2 of Theorem A.17. We bound the sum of the indices' conditional pseudoentropy gaps by combining their maximizing auxiliary variables into one variable in Definition D.3. Set

$$k_* = H(L) + \delta - \epsilon/2, \qquad \eta' = \epsilon/(4n), \qquad 2n\eta' = \epsilon/2.$$

Since $H(L)+\delta-\epsilon < k_*$, Definition D.3 gives an insertional decomposition $(L_0, \dots, L_n)$ for which every Z satisfying the length and (t,ϵ)-indistinguishability conditions in that definition satisfies

$$nH(Z \mid I, L_{I-1}) - \sum_{i=1}^{n} H(L_i \mid L_{i-1}) < k_* - H(L) = \delta - \epsilon/2.$$

Here I is uniform on $[n]$ and independent of the decomposition.

For each i, let $K_i = i(2^\ell - 1) + 1$ and let $\mathsf{Ext}_i(w, \cdot) : [K_i] \to \mathcal{E}_i(w)$ be the bijection in Lemma D.8. Define $A_i \in [K_i]$ by $L_i = \mathsf{Ext}_i(L_{i-1}, A_i)$. Since the map is a bijection for each fixed $L_{i-1} = w$,

$$H(A_i \mid L_{i-1}) = H(L_i \mid L_{i-1}).$$

Let $R(n, 2^\ell)$ be a fixed polynomial bounding the circuit size of each Ext_i and the overhead for composing it with a circuit whose index input is fixed to i, and put $t' = t + R(n, 2^\ell)$. Such a polynomial exists by the algorithms in Lemma D.8.

Let h_i be the maximum (t', η') conditional pseudoentropy of A_i given L_{i-1} over the alphabet $[K_i]$, and put

$$g_i = h_i - H(A_i \mid L_{i-1}) \geq 0.$$

Nonnegativity follows by using A_i itself as an auxiliary variable. We will show $\sum_i g_i < \delta - \epsilon/2$. By the compactness observation after Definition A.14, choose maximizing $A_i' \in [K_i]$ such that $H(A_i' \mid L_{i-1}) = h_i$ and (L_{i-1}, A_i') and (L_{i-1}, A_i) are (t', η') nonuniformly indistinguishable. Realize these pairs jointly with $(L_0, \dots, L_n)$ by setting

$$\begin{aligned}\Pr[A_1' = a_1, \dots, A_n' = a_n \mid L_0 = w_0, \dots, L_n = w_n] \\ = \prod_{i=1}^{n} \Pr[A_i' = a_i \mid L_{i-1} = w_{i-1}]\end{aligned}$$

for every tuple $(w_0, \dots, w_n)$ of positive probability, using the chosen pair distributions on the right. This preserves each distribution (L_{i-1}, A_i'). Put

$$Z_i = \mathsf{Ext}_i(L_{i-1}, A_i'), \qquad H(Z_i \mid L_{i-1}) = H(A_i' \mid L_{i-1}) = h_i,$$

choose I independently of this family and the decomposition, and set $Z = Z_I$.

Fix any Boolean circuit D of size at most t. For each fixed i, the circuit with inputs $w \in (\{0,1\}^\ell)^{i-1}$ and $a \in [K_i]$ given by

$$D_i(w, a) = D(i, \mathsf{Ext}_i(w, a), w)$$

has size at most $t + R(n, 2^\ell) = t'$. The defining indistinguishability of A_i' gives

$$\begin{aligned}&|\Pr[D(I, Z, L_{I-1}) = 1] - \Pr[D(I, L_I, L_{I-1}) = 1]| \\ &= \left| \frac{1}{n} \sum_{i=1}^{n} (\Pr[D_i(L_{i-1}, A_i') = 1] - \Pr[D_i(L_{i-1}, A_i) = 1]) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} |\Pr[D_i(L_{i-1}, A_i') = 1] - \Pr[D_i(L_{i-1}, A_i) = 1]| \\ &\leq \eta' \leq \epsilon.\end{aligned}$$

The family $(Z_1, \dots, Z_n)$ was fixed before D, and $Z \in (\{0,1\}^\ell)^i$ when $I = i$. Thus Z satisfies the length and indistinguishability conditions for the chosen decomposition, so

$$\begin{aligned}\sum_{i=1}^{n} g_i &= \sum_{i=1}^{n} h_i - \sum_{i=1}^{n} H(A_i \mid L_{i-1}) \\ &= \sum_{i=1}^{n} H(Z_i \mid L_{i-1}) - \sum_{i=1}^{n} H(L_i \mid L_{i-1}) \\ &= nH(Z \mid I, L_{I-1}) - \sum_{i=1}^{n} H(L_i \mid L_{i-1}) \\ &< k_* - H(L) = \delta - \epsilon/2.\end{aligned}$$

Apply item 2 of Theorem A.17 for each i with $X = L_{i-1}$, $B = A_i$, $m = m_i = (i-1)\ell \geq 0$, $q = K_i \geq 2$, circuit-size parameter t', gap bound g_i, and $\eta = \eta'$. It gives a randomized circuit $S_i : \{0,1\}^{m_i} \to [K_i]$ satisfying

$$|S_i| \leq p_0(t', m_i, K_i, 1/\eta'),$$
$$D_{\mathsf{KL}}\left((L_{i-1}, A_i) \parallel (L_{i-1}, S_i(L_{i-1}))\right) \leq g_i + 2\eta'.$$

Set $C_i(w) = \mathsf{Ext}_i(w, S_i(w))$, with fresh independent coins at each step. Since $S_i(w) \in [K_i]$, $C_i(w) \in \mathcal{E}_i(w)$ for every input $w \in (\{0,1\}^\ell)^{i-1}$. The bijection $(w, a) \mapsto (w, \mathsf{Ext}_i(w, a))$ gives

$$\begin{aligned} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) &= D_{\mathsf{KL}}\left((L_{i-1}, A_i) \parallel (L_{i-1}, S_i(L_{i-1}))\right) \\ &\leq g_i + 2\eta'. \end{aligned}$$

Hence

$$\begin{aligned} \sum_{i=1}^{n} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) &\leq \sum_{i=1}^{n} g_i + 2n\eta' \\ &< \delta - \epsilon/2 + \epsilon/2 = \delta. \end{aligned}$$

Dividing by n gives the stated average bound.

Finally, $m_i \leq n2^\ell$, $K_i \leq n2^\ell$, and $1/\eta' = 4n/\epsilon$. Since p_0 has nonnegative coefficients,

$$\begin{aligned} |C_i| &\leq |S_i| + R(n, 2^\ell) \\ &\leq p_0(t + R(n, 2^\ell), n2^\ell, n2^\ell, 4n/\epsilon) + R(n, 2^\ell). \end{aligned}$$

Choose the polynomial p in Theorem D.6 large enough to dominate this quantity at $(t, n, 2^\ell, 1/\epsilon)$ and the bound required in its first part. Then $B := \max_i |C_i| \leq p(t, n, 2^\ell, 1/\epsilon)$. □

Proof of Theorem D.6. For the first part, fix a sampler C satisfying the size bound.

Write $d = D_{\mathsf{KL}}\left(L \parallel C(1^n)\right)$. If $d = \infty$, the claim follows. Otherwise, Lemma D.9 gives an insertional decomposition $(L_0, \ldots, L_n)$ of L whose sum of conditional KL-divergences from C equals d. Applying Lemma D.10 to this decomposition gives

$$\begin{aligned} D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) &= \sum_{i=1}^{n} D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) \\ &> \delta - \sigma - n\eta. \end{aligned}$$

For the second part, take the decomposition and sampler from Lemma D.11. The sampler has the required size, and Lemma D.9 gives

$$D_{\mathsf{KL}}\left(L \parallel C(1^n)\right) \leq \sum_i D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right) \leq \delta.$$

□

E An Insertionally-Samplable PRG

Proposition E.1 (Insertionally samplable pseudorandom generators)**.** *Let $G : \{0,1\}^t \to \{0,1\}^{2t^2}$ be a polynomial-time pseudorandom generator secure against nonuniform polynomial-size circuits. There exists a polynomial-time pseudorandom generator $H : \{0,1\}^{3t^2} \to \{0,1\}^{4t^2}$ whose output distribution admits an exact polynomial-size stateless insertional bit sampler.*

Proof. Define H by the following process, with its seed consisting of the $3t^2$ independent uniform bits used below.

1. Append t^2 independent uniform bits, obtaining $X = X_1 \cdots X_{t^2}$. Partition X into t consecutive groups of t bits. Define s_j as the parity of the bits in group j, and put $s = s_1 \cdots s_t$.

2. For $a = 1, \ldots, 2t^2$, read X from the first t^2 positions, recompute s, and append $G(s)_a$. (Following this process, the resulting string is $X \| G(s)$.)

3. For $i = 1, \ldots, t^2$, sample fresh independent uniform bits b, r. Insert b immediately before X_i if $r = 0$, and immediately after it otherwise. Denote the resulting ordered pair by Z_i, and put $Z = Z_1 \| \cdots \| Z_{t^2}$. The final output is $Z \| G(s)$.

The process uses t^2 random bits in item 1 and $2t^2$ in item 3, and outputs $4t^2$ bits. It is stateless: throughout item 2, X occupies the first t^2 positions. After processing $i-1$ bits in item 3, the current string is

$$Z_1 \| \cdots \| Z_{i-1} \| X_i \cdots X_{t^2} \| G(s),$$

so X_i occupies position $2i - 1$. The current length determines the phase and index. Every transition therefore uses only the current string and fresh coins, and has polynomial-size circuits.

It remains to prove pseudorandomness. The pairs Z_i are independent and uniform over $\{00, 01, 10, 11\}$, so Z is uniform. Conditioned on any $Z = z$, the original bits X_i remain independent: an equal pair determines its original bit, whereas an unequal pair leaves its original bit uniform. Hence, if a group contains an unequal pair, its parity is uniform given $Z = z$.

The probability that all t pairs of a given group are equal is 2^{-t}. Except on an event of probability at most $t2^{-t}$, every group therefore has an unequal pair, making all seed bits independent and uniform given Z. Consequently,

$$\Delta\big((Z, s), (U_{2t^2}, U_t)\big) \le t2^{-t},$$

where the uniforms on the right are independent. Applying $(z, s) \mapsto z \| G(s)$ shows that $H(U_{3t^2})$ is statistically close to $U_{2t^2} \| G(U_t)$. By security of G, the latter is computationally indistinguishable from U_{4t^2}, proving security of H. □

F Separations in Sampling and Pseudoentropy

We separate insertional from sequential sampling over a binary alphabet in Section F.1, and general from insertional sampling over a larger alphabet in Section F.2. In Section F.3, we derive the corresponding separations between pseudoentropy notions.

F.1 Separating Insertional Samplers from Sequential Samplers

In this section formalize the intuition described in (Wu, 2023) who described a distribution that was hard for sequential samplers to derive. Moreover, we show that the same intuition does not apply to insertional samplers, thereby showing that insertional samplers cover a larger class of distributions compared to their sequential counterparts.

Proposition F.1 (Insertional sampling is stronger than sequential sampling)**.** *Let $f\colon \{0,1\}^n \to \{0,1\}^n$ be a one-way function secure against polynomial-size circuits. There exists a distribution L over $\{0,1\}^{2n}$ such that:*

1. *Every stateless sequential sampler of size $\mathsf{poly}(n)$ satisfies $D_{\mathsf{KL}}\left(L \parallel C(1^{2n})\right) > \log(n)$, and*

2. *There is a $\mathsf{poly}(n)$-size insertional sampler that samples L.*

Proof. We set L to be the distribution $L = (f(X), X)$ where X is uniform over $\{0,1\}^n$.

We now prove the first item. Suppose towards contradiction that for some constant $c > 0$ and infinitely many n, a size-n^c stateless sequential sampler C satisfies $D_{\mathsf{KL}}\left(L \parallel C(1^{2n})\right) \le \log n$. We show that this invalidates security of the one-way function. Given $y = f(x)$ the inverter I runs C sequentially on prefix y and outputs the n-bit suffix generated by C.

For $X \leftarrow U_n$, put $Y = f(X)$ and write $(V, W) \leftarrow C(1^{2n})$, where $V, W \in \{0,1\}^n$. Since C is stateless, $W \mid_{V=y}$ is distributed as $I(y)$ for every $y \in \mathsf{supp}(V)$. The assumed finite KL-divergence

ensures that $\mathsf{supp}(Y) \subseteq \mathsf{supp}(V)$. Applying the KL chain rule (Fact A.7) to both pairs gives

$$\begin{aligned} D_{\mathsf{KL}}\left((Y,X) \parallel (Y,I(Y))\right) &= \mathop{\mathbb{E}}_{y \leftarrow Y}\left[D_{\mathsf{KL}}\left(X\,|_{Y=y} \parallel I(y)\right)\right] \\ &= D_{\mathsf{KL}}\left(L \parallel C(1^{2n})\right) - D_{\mathsf{KL}}\left(Y \parallel V\right) \le \log n, \end{aligned}$$

where the inequality uses nonnegativity of KL-divergence.

Let $p = \Pr[f(I(Y)) = Y]$ be the inverter's success probability. The test $T(y, x') = \mathbb{1}[f(x') = y]$ maps (Y, X) to the constant 1 and $(Y, I(Y))$ to a Bernoulli random variable with parameter p. Thus data processing (Fact A.9) gives

$$\begin{aligned} \log(1/p) &= D_{\mathsf{KL}}\left(\mathrm{Ber}(1) \parallel \mathrm{Ber}(p)\right) \\ &\le D_{\mathsf{KL}}\left((Y,X) \parallel (Y,I(Y))\right) \le \log n. \end{aligned}$$

Hence the polynomial-size inverter I succeeds with probability $p \ge 1/n$ for infinitely many n, contradicting one-wayness.

We now show the second item by describing an insertional sampler for L:

1. For the first n steps, append independent uniform bits, obtaining $x \leftarrow U_n$.
2. At step $n + j$, for $j \in [n]$, read x from the input, compute $f(x)$ and output $(j, f(x)_j)$ (i.e., place the j-th bit of $f(x)$ at index j.

That this insertional sampler samples L exactly and is efficient is immediate by construction.

□

F.2 Separating General Samplers from Insertional Samplers

We construct an efficiently samplable distribution that is hard for stateless insertional samplers.

Theorem F.2 (General sampling is stronger than insertional sampling)**.** *Let $(f_\lambda)_{\lambda \ge 1}$ be a length-preserving one-way function family with $2^{\gamma\lambda^\alpha}$ security against circuits, for constants $\gamma > 0$ and $1/2 < \alpha < 1$. For $t \ge 2$, let $\ell = 2t + \lceil \log(2t) \rceil$. There exists a distribution $L = L_t$ over $(\{0,1\}^\ell)^{2t}$ such that:*

1. *For every constant $c > 0$ and all sufficiently large t, every stateless insertional sampler C of size at most 2^{ct} satisfies $D_{\mathsf{KL}}\left(L \parallel C(1^{2t})\right) \ge \gamma t^{2\alpha} - 2t$.*
2. *There is a $\mathsf{poly}(t)$-size randomized circuit that samples L exactly.*

Proof. We first define L. Let $q = 2^t$, fix distinct $\beta_1, \ldots, \beta_{2t} \in \mathbb{F}_q$, and identify $\mathbb{F}_q^t$ with $\{0,1\}^{t^2}$. For $x, r, y \in \mathbb{F}_q^t$, put $P_{x,r}(z) = \sum_{j<t} r_j z^j + \sum_{j<t} x_j z^{t+j}$ and $Q_y(z) = \sum_{j<t} y_j z^j$. The distribution L is sampled as follows:

1. Sample independent uniform $X, R \leftarrow \mathbb{F}_q^t$ and compute $Y = f_{t^2}(X)$.
2. Output $((i, P_{X,R}(\beta_i), Q_Y(\beta_i)))_{i=1}^{2t}$, treating each triple as one ℓ-bit token.

The evaluations of $P_{X,R}$ form a ramp secret-sharing of X: any fixed t shares are uniform and independent of X, since the Vandermonde matrix acting on R is invertible, while all $2t$ shares recover X, R by interpolation. Intermediate numbers of shares may reveal partial information about X. In contrast, any t evaluations of Q_Y determine Y. Thus we can sample any fixed t tokens given only the image, but completing them to a valid output reveals a preimage.

Observe that the description of the distribution defines a circuit for sampling from it. This sampler runs in $\mathsf{poly}(t)$ time, proving the second item. Its alphabet is $[2t] \times \mathbb{F}_q^2$, which has size $2t\,2^{2t}$.

We now prove the first item. Suppose towards contradiction that for some constant $c > 0$ and infinitely many t, a stateless insertional sampler C of size at most 2^{ct} has KL error $d := D_{\mathsf{KL}}\left(L \parallel C(1^{2t})\right)$ smaller than $\gamma t^{2\alpha} - 2t$. We show that this invalidates security of the one-way function.

Let $\mathcal{S} = \mathsf{supp}(L)$ and observe that membership in $\mathcal{S}$ can be efficiently tested by interpolating to derive X and then checking whether the third entries in the output are consistent with $f(X)$. We now describe the inverter. Let $b = \lceil \log \binom{2t}{t} \rceil \le 2t$. Given $y = f_{t^2}(x)$, the inverter I proceeds as follows:

1. Choose a t-element subset $J \subseteq [2t]$ by reducing a uniform b-bit integer modulo $\binom{2t}{t}$ and taking the corresponding subset in lexicographic order.
2. Sample independent uniform $u_j \leftarrow \mathbb{F}_q$ for $j \in J$. Set $W_t = ((j, u_j, Q_y(\beta_j)))_{j \in J}$ in increasing tag order.
3. For $i = t+1, \dots, 2t$, sample $W_i \leftarrow C_i(W_{i-1})$ using fresh coins.
4. If $W_{2t} \in \mathcal{S}$, output the preimage recovered by interpolating the entries that encode x; otherwise output $\perp$.

On acceptance, the preserved t evaluations of Q_y agree with those of $Q_{f_{t^2}(x')}$ for the recovered x'. Both polynomials have degree less than t, so $f_{t^2}(x') = y$. The inverter has size $2^{O(t)}$: it makes t calls to C and otherwise uses $\mathsf{poly}(t)$ computation.

It remains to lower-bound its success probability. Below, the strings $W_t, \dots, W_{2t}$ are generated on a random challenge $Y = f_{t^2}(X)$ with $X \leftarrow U_{t^2}$. By Lemma D.9, there is a Markov insertional decomposition $(L_0, \dots, L_{2t})$ of L such that, writing $d_i = D_{\mathsf{KL}}\left((L_{i-1}, L_i) \parallel (L_{i-1}, C_i(L_{i-1}))\right)$, we have $\sum_{i=1}^{2t} d_i = d$.

For a fixed set J of t tags, write $L|_J$ for the restriction of L to those tags. The secret-sharing property implies that W_t, conditioned on choosing J, is distributed as $L|_J$. Moreover, each J is chosen with probability at least 2^{-b}. If w has tag set J, the event $L_t = w$ implies $L|_J = w$. Consequently, for every $w \in \mathsf{supp}(L_t)$,

$$\Pr[L_t = w] \le \Pr[L|_J = w] \le 2^b \Pr[W_t = w],$$

and hence $D_{\mathsf{KL}}\left(L_t \parallel W_t\right) \le b$. This comparison allows the indices in L_t to depend on the final string; it uses secrecy only for each fixed J.

Since C is stateless, the inverter continues from W_t using only the current string. The KL chain rule (Fact A.7) and nonnegativity of the d_i therefore give

$$D_{\mathsf{KL}}\left((L_t, \dots, L_{2t}) \parallel (W_t, \dots, W_{2t})\right) = D_{\mathsf{KL}}\left(L_t \parallel W_t\right) + \sum_{i=t+1}^{2t} d_i \le b + d.$$

Let $p = \Pr[W_{2t} \in \mathcal{S}]$ be the inverter's success probability. The support test accepts L_{2t} with probability one and W_{2t} with probability p. Thus data processing (Fact A.9) gives

$$\log(1/p) = D_{\mathsf{KL}}\left(\mathrm{Ber}(1) \parallel \mathrm{Ber}(p)\right) \le b + d < \gamma t^{2\alpha}.$$

Hence $p > 2^{-\gamma t^{2\alpha}}$. Since $2\alpha > 1$, the inverter's size $2^{O(t)}$ is at most $2^{\gamma t^{2\alpha}}$ for all sufficiently large t. This contradicts the assumed security at input length $\lambda = t^2$ for infinitely many t, proving the second item. □

F.3 Separating Pseudoentropy Notions

We now combine the sampler separations with Theorems B.1 and D.6 to separate the corresponding pseudoentropy notions. For each notion, the pseudoentropy gap is its pseudoentropy minus the Shannon entropy of the distribution.

Corollary F.3 (Separating insertional from next-block pseudoentropy gaps)**.** *Let $(f_n)_{n \ge 1}$ be a length-preserving one-way function family secure against polynomial-size circuits. For $X \leftarrow U_n$, let $L = (f_n(X), X)$, viewed as $2n$ binary tokens. For every sufficiently large constant c and all sufficiently large n:*

1. *The (n^c, n^{-c}) insertional pseudoentropy gap of L is at most $1/n$.*
2. *The (n^c, n^{-c}) next-block pseudoentropy gap of L is at least $\frac{1}{2} \log n$.*

Proof. Since L includes the uniform preimage, $H(L) = n$. For the first item, use the fixed insertion order from the proof of Proposition F.1, and let $(L_0, \dots, L_{2n})$ be its intermediate strings. Write $A_i(w)$ for the possible successors of w: for $i \le n$, these are $w0, w1$; for $i > n$, there is only the prescribed insertion of an image bit. A circuit of size $T_0(n)$ checks whether $z \in A_i(w)$, for some polynomial T_0. Fix $c \ge 4$ large enough that $T_0(n) \le n^c$ for all sufficiently large n, and set $\varepsilon = n^{-c}$.

Let I be uniform on $[2n]$ and let Z be an auxiliary variable satisfying the (n^c, ε) indistinguishability condition in Definition D.3 for this decomposition. Put $a = \Pr[Z \notin A_I(L_{I-1})]$. The test above gives $a \le \varepsilon$. Conditioning on whether this event occurs gives

$$\begin{aligned} H(Z \mid I, L_{I-1}) &\le h(a) + \mathbb{E}[\log |A_I(L_{I-1})|] + 2na \\ &\le h(\varepsilon) + \tfrac{1}{2} + 2n\varepsilon. \end{aligned}$$

Since $H(L) = n$, the insertional pseudoentropy gap is at most $2nh(\varepsilon) + 4n^2\varepsilon$. Since $h(\varepsilon) = O(\varepsilon \log(1/\varepsilon))$, this is $O(n^{2-c}) \le 1/n$ for all sufficiently large n, proving the first item.

For the second item, suppose that for some $c > 0$ and infinitely many n, the (n^c, n^{-c}) next-block pseudoentropy gap is less than the claimed bound. For all sufficiently large such n, $\frac{1}{2}\log n \le \log n - n^{-c}$. Part 2 of Theorem B.1, with $\delta = \log n$ and $\epsilon = n^{-c}$, gives a stateless sequential sampler of size $\mathsf{poly}(n^c, 2n, 2, n^c) = \mathsf{poly}(n)$ and KL error at most $\log n$, contradicting Proposition F.1. □

Corollary F.4 (Separating insertional from ordinary pseudoentropy gaps)**.** *Let $(f_\lambda)_{\lambda \ge 1}$ be a length-preserving one-way function family with $2^{\gamma\lambda^\alpha}$ security against circuits, for constants $\gamma > 0$ and $1/2 < \alpha < 1$. For $t \ge 2$, let $\ell = 2t + \lceil \log(2t) \rceil$. There is a distribution $L = L_t$ over $(\{0,1\}^\ell)^{2t}$, samplable in $\mathsf{poly}(t)$ time, such that for every sufficiently large constant c and all sufficiently large t:*

1. *The ordinary (t^c, t^{-c}) pseudoentropy gap of L is at most $1/t$.*
2. *The (t^c, t^{-c}) insertional pseudoentropy gap of L is at least $\frac{\gamma}{2}t^{2\alpha}$.*

Proof. Use the distribution constructed in Theorem F.2. Every output determines the independent uniform X, R, so L is uniform on its support $\mathcal{S}$ of size 2^{2t^2} and $H(L) = 2t^2$. Let $T_0(t)$ be a polynomial bound on the size of the support test in that proof. Fix $c \ge 4$ large enough that $T_0(t) \le t^c$ for all sufficiently large t, and set $\varepsilon = t^{-c}$. If V is (t^c, ε)-indistinguishable from L, then $a = \Pr[V \notin \mathcal{S}] \le \varepsilon$. Conditioning on membership in $\mathcal{S}$ gives

$$H(V) \le h(a) + (1-a)2t^2 + a(2t\ell) \le H(L) + h(\varepsilon) + 2t\ell\varepsilon,$$

so the ordinary pseudoentropy gap is at most $h(\varepsilon) + 2t\ell\varepsilon = O(t^{2-c}) \le 1/t$ for all sufficiently large t, proving the first item.

For the second item, suppose towards contradiction that for some $c > 0$ and infinitely many t, the (t^c, t^{-c}) insertional pseudoentropy gap is less than $\frac{\gamma}{2}t^{2\alpha}$. Set $\delta = \frac{\gamma}{2}t^{2\alpha} + t^{-c}$, so the insertional pseudoentropy is at most $H(L) + \delta - t^{-c}$. Part 2 of Theorem D.6 gives a stateless insertional sampler of size $\mathsf{poly}(t^c, 2t, 2^\ell, t^c) = 2^{O(t)}$ and KL error at most δ. Since $2\alpha > 1$, for all sufficiently large t we have $\delta < \gamma t^{2\alpha} - 2t$, contradicting Theorem F.2. □

G Cryptographic Consequences

Our results show that small-state sequential samplers cannot do many standard cryptographic tasks. We give two examples: (1) a small-state sequential sampler cannot sample a pseudorandom generator distribution, and (2) a small-state sequential sampler cannot sampler fresh public-key ciphertexts even given the public key. This shows that, in order to compute cryptographic primitives, autoregressive models require either external mechanisms or a large state. On the flip-side, we show that sequential samplers cannot invoke public-key steganography. Thus, agents with a small state cannot covertly communicate with each other. This may be useful for security and alignment, in cases where an autoregressive agent is not trusted.

In order to describe consequences of our results to sequential generation of cryptographic primitives, we first derive a general consequence of Theorems B.1 and C.4 for negligible sampling error, allowing logarithmic private state. We use n as the asymptotic parameter throughout this subsection.

Corollary G.1 (General consequence for samplers with logarithmic state)**.** *Let (W, Y) be jointly distributed and indexed by n. Assume:*

1. *W has polynomial length, and Y consists of $\mathsf{poly}(n)$ items over a $\mathsf{poly}(n)$-size alphabet.*
2. *Y has a noticeable nonuniform next-block pseudoentropy gap given W.*

Then no sequential sampler C with $\mathsf{poly}(n)$-size circuits and $O(\log n)$ private bits satisfies

$$D_{\mathsf{KL}}\left((W, Y) \parallel (W, C(W))\right) = \mathsf{negl}(n),$$

even given W and the entire visible prefix at every step.

Proof. Suppose towards contradiction that such a sampler C exists, and write $k = O(\log n)$ for its state length. Let $N \geq 1$ be the number of tokens of Y, each of length ℓ. Put $P = (W, Y)$, $Q = (W, C(W))$, and $e = D_{\mathsf{KL}}\left(P \parallel Q\right) = \mathsf{negl}(n)$. Choose $d > 0$ so that the pseudoentropy gap is at least n^{-d} for all sufficiently large n, and set $\rho = n^{-d}/8$. We remove the private state, control the resulting error, and then apply the stateless characterization.

1. *Remove the private state.* Apply Theorem C.4 with accuracy ρ. Its construction allows visible input W: use the same w in every simulated invocation of C, and average the KL bound over W. The resulting stateless sampler D has polynomial size, since $2^k, 2^\ell, 1/\rho, N, |W|$ are polynomially bounded. Writing $R = (W, D(W))$, we have $D_{\mathsf{KL}}\left(Q \parallel R\right) \leq \rho$.

2. *Compare with the target distribution.* Both C and D use at most $r = \mathsf{poly}(n)$ fair random bits over the entire run. Thus every positive conditional output probability, given $W = w$, is at least 2^{-r}. Finite KL-divergence gives $\mathsf{supp}(P) \subseteq \mathsf{supp}(Q) \subseteq \mathsf{supp}(R)$. On $\mathsf{supp}(Q)$, put $F(w, y) = \log(\Pr[C(w) = y] / \Pr[D(w) = y])$, and set $F = 0$ elsewhere. Then $|F| \leq r$. By Fact A.5, $\nu := \Delta(P, Q) = \mathsf{negl}(n)$. Since all three distributions have the same W marginal,

$$\begin{aligned} D_{\mathsf{KL}}\left(P \parallel R\right) &= e + D_{\mathsf{KL}}\left(Q \parallel R\right) + \mathop{\mathbb{E}}_{P} F - \mathop{\mathbb{E}}_{Q} F \\ &\leq e + \rho + 2r\nu = \rho + \mathsf{negl}(n) \leq 2\rho \end{aligned}$$

   for all sufficiently large n.

3. *Apply the pseudoentropy lower bound.* Part 1 of Theorem B.1 allows visible input W: replace $(I, Y_{<I})$ by $(W, I, Y_{<I})$, condition the entropies on W, and include $|W|$ in the polynomial size bound. Set $\delta = 4\rho$, $\sigma = \rho$, $\eta = \rho/N$, $t = n^c$, and $\epsilon = n^{-c}$. The denominator in the size bound is polynomial in n independently of c, while the numerator is $\min\{n^{cc_0}, \rho n^c/N\}$. Choose c large enough to cover the size of D and ensure $n^{-c} \leq \rho$. The pseudoentropy hypothesis then gives (n^c, n^{-c}) next-block pseudoentropy given W at least $H(Y \mid W) + 8\rho - n^{-c} \geq H(Y \mid W) + 4\rho$. The theorem yields $D_{\mathsf{KL}}\left(P \parallel R\right) > 4\rho - \rho - N(\rho/N) = 2\rho$, contradicting the preceding bound.

□

We now describe several corollaries for various cryptographic consequences of the above corollary.

Corollary G.2 (Pseudorandom generation)**.** *Let $G : \{0,1\}^{k(n)} \to \{0,1\}^n$ be a pseudorandom generator secure against nonuniform polynomial-size circuits. No sequential bit sampler C with $\mathsf{poly}(n)$-size circuits and $O(\log n)$ private bits satisfies*

$$D_{\mathsf{KL}}\left(G(U_{k(n)}) \parallel C(1^n)\right) = \mathsf{negl}(n).$$

The same holds for any apriori fixed ordering, rather than just left-to-right.

Corollary G.3 (Public-key encryption)**.** *Let a polynomial-time public-key encryption scheme have security parameter n. Assume:*

1. *IND-CPA security against nonuniform polynomial-size circuits and negligible correctness error.*
2. *Ciphertexts consist of a fixed polynomially bounded number of $O(\log n)$-bit tokens.*

For any two distinct equal-length messages of polynomial length, their ciphertext distributions cannot both be sampled with negligible KL-divergence by sequential samplers with $\mathsf{poly}(n)$*-size circuits and* $O(\log n)$ *private bits. The public key, message, and visible prefix are available at every step; divergence is measured jointly with the public key.*

Corollary G.4 (Public-key steganography)**.** *Let a public-key steganography scheme run in polynomial time, with security parameter* n*. Fix two distinct messages of polynomial length. Assume:*

1. *Negligible decoding error; both encodings consist of the same fixed polynomially bounded number of* $O(\log n)$*-bit tokens.*
2. *Both encodings are computationally indistinguishable from the same cover distribution, jointly with the public key, against nonuniform polynomial-size circuits.*

The two encoding distributions cannot both be sampled with negligible KL-divergence by sequential samplers with $\mathsf{poly}(n)$*-size circuits and* $O(\log n)$ *private bits. The public key, message, and visible prefix are available at every step; divergence is measured jointly with the public key.*

Proof of the three corollaries. For the PRG, we describe it for left-to-right ordering, but the same argument holds for any fixed known ordering up to applying permutation. Put $Y = G(U_{k(n)})$. For uniform $I \in [n]$ and an independent uniform bit B, the distributions $(I, Y_{<I}, Y_I)$ and $(I, Y_{<I}, B)$ are computationally indistinguishable: replacing Y by U_n on both sides makes them identical. Thus Y has next-bit pseudoentropy n, whereas $H(Y) \le k(n) < n$. Corollary G.1 applies with constant visible input.

For the remaining claims, let $P_b = (W, Y^b)$ be the two encoding distributions, including their common public key or setup W. Write N for the number of tokens and $u = 2^\ell$ for the alphabet size, where ℓ is the token length. They are computationally indistinguishable and satisfy $\Delta(P_0, P_1) \ge 1 - \mathsf{negl}(n)$. For encryption and steganography, correctness gives the distance bound: appending the secret key preserves distance, since it is independent of the encoding given the public key. Steganographic security gives indistinguishability through the common cover.

Suppose towards contradiction that both samplers exist. We distinguish P_0 from P_1 in three steps.

1. *Remove the private state.* Put $\alpha = 1/(100N)$. Apply Theorem C.4 with KL error $\alpha^2/4$, keeping W visible. By Fact A.5 and the assumed negligible sampling error, the resulting polynomial-size stateless samplers D^b satisfy $\Delta(P_b, R_b) \le \alpha$, where $R_b = (W, D^b(W))$. Hence $\Delta(R_0, R_1) \ge 1/3$ for all sufficiently large n.

2. *Find a differing transition.* Replacing the circuits of D^0 by those of D^1 one at a time gives an index i such that, for $X = (W, D^0(W)_{<i})$ and $A_b \leftarrow D_i^b(X)$ with fresh coins, $d := \Delta((X, A_0), (X, A_1)) \ge 1/(3N)$. Each hybrid uses the message-0 prefix and a common message-1 suffix, which cannot increase distance. Yet every polynomial-size test distinguishes these pairs with advantage at most $4\alpha + \mathsf{negl}(n)$: replacing the R_0 prefix by the R_1 prefix through P_0, P_1 before invoking D_i^1 costs $2\alpha + \mathsf{negl}(n)$, and comparing the two honest prefixes including token i costs the same amount.

3. *Test the transition probabilities.* On input (x, a), draw $M = \lceil 576uN^2 \rceil$ samples from each $D_i^b(x)$, and accept if a occurs more often in the first batch. Write p_b for the true token distribution and $\widehat{p}_b$ for its empirical distribution. By Cauchy–Schwarz, $\mathbb{E}\,\|\widehat{p}_b - p_b\|_1 \le \sqrt{u/M} \le 1/(24N)$. Choosing the larger empirical probability loses at most $\|\widehat{p}_0 - p_0\|_1 + \|\widehat{p}_1 - p_1\|_1$ relative to the optimal token test. Averaging over X, our test therefore has advantage at least $d - 1/(12N) \ge 1/(4N)$, contradicting the preceding upper bound.

The index i is nonuniform advice, and all circuits have polynomial size. The test's random coins can be fixed by averaging. □