EXACT PROBABILITY POWERING FROM SAMPLES ALONE AT THE COLLISION SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Temperature sampling normally needs the probability distribution, yet many systems, from simulators to language models reasoning toward a final answer, provide only *samples*. We ask how to sample exactly at temperature $1/k$ from samples alone, drawing each outcome with probability proportional to p_i^k, the kth power of its probability, for a fixed integer $k \geq 2$ (probability powering). Known approaches face a trade-off between speed and exactness. Classical rejection sampling, which keeps an outcome only when k fresh samples agree, is exact but slow: with n equally likely outcomes, it needs $\Theta(n^{k-1})$ samples. The first-collision shortcut, which returns the first outcome to appear k times among all samples drawn so far (a k-fold collision), is fast—needing only $\Theta(n^{1-1/k})$ samples—but biased. Refined collision-based methods reduce this bias but remain approximate. In this work, we show that this trade-off is unnecessary. We define the *collision map*, a random update built from one collision and the other outcomes seen before it, and prove that it preserves the target distribution. Our sampler, **Collision-map CFTP**, composes these maps backward in time (coupling from the past) *without knowing* the probabilities or which outcomes are possible. It returns an *exact* output after a bounded expected number of collisions, needing only $\Theta(n^{1-1/k})$ samples instead of rejection's $\Theta(n^{k-1})$. Conversely, every sampler that is exact for every distribution must see its output k times before returning it, so none can beat this method on *any* input by more than a factor depending only on k (instance optimality). *Exactness thus costs only a constant factor over the biased shortcut*, e.g., for squaring, we prove that this factor is below 11 on every input and tends to $8/3$ on large uniform inputs. Simulations match the predicted costs and output distributions. At temperature $1/2$ over a million equally likely outcomes, our Collision-map CFTP uses about 3,400 samples per output on average, whereas rejection needs about 2,100,000. All theorems are formally verified in Lean 4.

1 INTRODUCTION

Lowering the temperature of a distribution to $1/k$ reweights each outcome of probability p_i in proportion to p_i^k. The usual logit rescaling requires access to probabilities (Holtzman et al., 2020), but many generators and simulators provide only samples. Even with token probabilities, the probability of a final answer is an intractable sum over the reasoning paths that reach it (Dohan et al., 2022; Kuhn et al., 2023). A fixed rule extracting the answer from each generation nevertheless gives samples of that answer (Wang et al., 2023; Chen et al., 2023). *Using samples alone, how can we draw exactly from the temperature-$1/k$ distribution, and what does it cost?*

We study this question for a fixed integer $k \geq 2$ and an unknown discrete distribution p. We call the outcome of a sample its *label*, for example the extracted final answer, and the sampler can only test whether two labels are equal. The target of *probability powering* is the *powered distribution*, the conditional law of the common label of k independent samples $X_1, \ldots, X_k$ given that they agree:

$$\mathbb{P}(X_1 = i \mid X_1 = \cdots = X_k) = \frac{p_i^k}{\sum_j p_j^k} =: q_k(i).$$

It remains random but favors labels on which independent samples are likely to agree: an odds ratio of two becomes 2^k. Token-by-token temperature changes do not give even the powered distribution

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.

of full sequences (Karan & Du, 2026, Proposition 1), and neither they nor majority voting over a finite batch generally gives q_k. An exact sample needs no tuning parameter whose adequate value depends on the unknown p, keeps downstream Monte Carlo averages unbiased, and provides a ground truth for approximate sharpening.

The agreement formula suggests the classical exact method: draw k samples, return their common label if they all agree, and otherwise start again. This is the basic rejection approach to Bernoulli factories (Morina et al., 2022), and the continuous autoregressive language model CALM uses it for exact temperature sampling (Shao et al., 2025). Its expected cost is of order $1/\sum_j p_j^k$, that is, n^{k-1} on n equally likely labels, even for *early-abort rejection*, which abandons each group at its first mismatch. Matches anywhere in the stream arrive far sooner. By the generalized birthday problem (Flajolet et al., 1992; Holst, 1995; Camarri & Pitman, 2000), some label appears k times (a *k-fold collision*) after order $n^{1-1/k}$ samples (Theorem 1). However, the first label to do so, the *winner*, generally does not have law q_k (Section 4.1). Methods that exploit collisions have been approximate. For lattice algorithms, Aggarwal et al. (2015) construct an approximate sampler for the squared distribution q_2. For temperature sampling, CALM also gives a more sample-efficient batch method that weights each label by its number of matching k-subsets in a large batch. The method is consistent but biased for every finite batch (Shao et al., 2025). Must exactness give up the speed of collisions?

Exactness at the collision scale. We show that it need not, and that no *universal exact* sampler, one exact for every input distribution, can beat the collision speed. Let J_k count samples through the first time any label has appeared k times, and write $\|p\|_k = (\sum_i p_i^k)^{1/k}$. Our sampler returns an exact sample from q_k using at most $C_k\,\mathbb{E}_p J_k$ samples on average, where C_k depends only on k. Conversely, every universal exact sampler must have seen its returned label k times, so it uses at least $\mathbb{E}_p J_k$ samples. Since $\mathbb{E}_p J_k$ is within factors depending only on k of $1/\|p\|_k$, these bounds determine the optimal expected cost on every input up to such a factor (*instance optimality*). In terms of the Rényi entropy $H_k(p) = \frac{1}{1-k}\log\sum_i p_i^k$ (Rényi, 1961), exact powering costs order $e^{(1-1/k)H_k(p)}$ samples, whereas rejection costs order $e^{(k-1)H_k(p)}$. On n equally likely labels, $H_k(p) = \log n$. The sampler needs no probabilities and no knowledge of the support, which may be countably infinite.

Matching bounds. Both bounds are governed by k-fold repetition. A universal exact sampler cannot return a label seen fewer than k times: doing so would give it too much output probability when it is rare (Section 3). For the upper bound, we retain both the winner and the other labels seen in each *collision block*, the samples through a first k-fold collision. They define a random *collision map* preserving q_k. Coupling from the past (CFTP, Propp & Wilson, 1996) composes independent maps until all starting labels give the same output (*coalescence*), which is then exact. The expected cost is $\mathbb{E}_p J_k$ times the expected number of blocks. To bound the latter independently of p, we analyze the maps composed in the order drawn: after k blocks, every possible output other than at most $k-1$ winners must have appeared in all the blocks. The same power sum that controls block lengths bounds the expected number of these shared labels; each later block merges any two remaining outputs with probability bounded below in terms of k (Section 5.2).

Contributions.

- **Exact powering at the collision scale.** Collision-map CFTP (Algorithm 1) needs only IID samples and label comparisons, returns exactly q_k on any countable support, and uses at most $C_k\mathbb{E}_p J_k \le kC_k/\|p\|_k$ samples in expectation, with $C_2 < 11$ (Theorem 1). On n equally likely labels this is of order $n^{1-1/k}$, whereas rejection needs order n^{k-1}.
- **A matching lower bound on every input.** Any universal exact sampler draws at least J_k samples almost surely (Theorem 2). The biased first-collision shortcut pays exactly this lower bound, and exactness adds a factor depending only on k.
- **Exact sampling without a known state space.** A collision map gives a transition that is reversible for q_k. The composed map is stored as a default output plus a finite dictionary of exceptions, so an empty dictionary certifies coalescence. Intersecting k independent blocks bounds the expected number of possible outputs independently of the support (Sections 4 and 5). These tools may transfer to other CFTP constructions on unknown countable spaces.

- **Exact constants, numerics, and verification.** On large uniform inputs the mean number of blocks tends to an explicit constant between k and $k+1$ ($8/3$ for squaring), and simulations match the predicted cost and output law (Section 6). All theorems, lemmas, and propositions are formally verified (Appendix D) in Lean 4 with Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020).

Relation to prior work. Bernoulli factories turn flips of a coin with unknown bias into flips with a transformed bias (Keane & O'Brien, 1994; Nacu & Peres, 2005; Łatuszyński et al., 2011; Huber, 2016; 2017), with multivariate extensions for exact reductions and combinatorial sampling (Dughmi et al., 2017; Niazadeh et al., 2023). The closest framework, Dice Enterprise (Morina et al., 2022), constructs a reversible Markov chain for rational transformations, including integer powers, on a known finite alphabet and samples its stationary law by CFTP. We drive the chain with collision blocks instead, allowing unknown countable support, and establish matching sample-cost bounds on every input. Marginal sharpening targets the same answer distribution using token probabilities conditioned on sampled reasoning traces (Arzhantsev et al., 2026); full-sequence power samplers also use likelihoods (Karan & Du, 2026; Azizi et al., 2026; Ji et al., 2026; Zhou et al., 2026). Our sampler requires only IID outputs. Appendix C.1 compares other exact-sampling frameworks and the related collision scale in entropy estimation.

2 Problem formulation and main results

To compare algorithms with the same information, we first specify their access to the unknown distribution. We consider a fixed countable label universe $\mathcal{X}$ with at least two possible labels. A sampling oracle returns independent and identically distributed (IID) samples from one fixed, unknown distribution p on $\mathcal{X}$, whose positive support (the labels with $p_i > 0$) may be finite or countably infinite. This models a generator whose outputs can be compared but whose probabilities cannot be queried. For a fixed integer $k \geq 2$, define

$$S_k = \sum_{i \in \mathcal{X}} p_i^k, \qquad \|p\|_k = S_k^{1/k}, \qquad q_k(i) = \frac{p_i^k}{S_k}.$$

The sampler receives k and can compare labels. It receives no probabilities, support enumeration, support-size bound, or other p-dependent information. A *universal exact sampler* is one algorithm that terminates almost surely and returns q_k for every distribution in this model, which is what not knowing p means. We measure the cost of one output by the total number of samples drawn, including discarded ones, since each sample is one call to the generator. The *first k-fold collision time* J_k counts samples through the first occurrence of any label for the kth time, including the final one. Write $\mu_k = \mathbb{E}_p[J_k]$. In O_k and Θ_k, the hidden constants depend on k alone, not on the size of the support.

2.1 Complexity guarantees

Algorithm 1, constructed in Section 4, repeatedly draws collision blocks and uses each block to define a random map from labels to labels.

Theorem 1 (Exact sampling with cost at the collision scale)**.** *There is a finite constant C_k depending only on k such that, for every distribution in the stated oracle model, Algorithm 1 terminates almost surely and returns exactly q_k, and its total number of samples N satisfies*

$$\mathbb{E}_p[N] \leq C_k \mu_k \leq \frac{kC_k}{\|p\|_k}, \qquad \mu_k \geq \frac{(k!)^{1/k}\Gamma(1+1/k)}{\|p\|_k}.$$

For $k = 2$, one may take $C_2 = 691/64 < 11$, hence $\mathbb{E}_p[N] < 11\mu_2 \leq 22/\|p\|_2$. The expected number of dictionary operations and the expected peak number of stored label entries are $O_k(1/\|p\|_k)$.

Theorem 2 (An instancewise lower bound for universal exact samplers)**.** *Let $\mathcal{B}$ be a universal exact sampler. For every fixed p, its returned label has already occurred at least k times almost surely. Consequently, its sample count $N_{\mathcal{B}}$ satisfies $N_{\mathcal{B}} \geq J_k$ almost surely and $\mathbb{E}_p[N_{\mathcal{B}}] \geq \mu_k$. Algorithm 1 is therefore within a factor C_k of every such algorithm's expected cost on each individual p.*

Necessity and sufficiency on each input. For our sample count N, the two theorems give

$$\mu_k \le \inf_{\mathcal{B} \text{ universal exact}} \mathbb{E}_p[N_{\mathcal{B}}] \le \mathbb{E}_p[N] \le C_k \mu_k.$$

The sampler is therefore optimal on every input, not only on the hardest one, up to a factor depending only on k. Support size alone does not describe this complexity, since distributions on the same support can have very different $\|p\|_k$. The sampler adapts through the lengths of its collision blocks without estimating this norm. The gain over rejection holds input by input: rejection costs between $2/S_k$ and $k/S_k = k/\|p\|_k^k$ samples, so our cost is within a factor depending only on k of the kth root of the rejection cost.[1]

For this construction the factor $\mathbb{E}_p[N]/\mu_k$ tends to a value between k and $k+1$ on large uniform inputs (Section 6) but equals 2^{k-1} on a fair coin (Appendix B.1), so very low temperatures remain costly. The comparison class is essential: a sampler promised a uniform input could return its first sample, because powering leaves a uniform distribution unchanged.

3 Every universal exact sampler must wait for a collision

The lower bound does not depend on our construction, so we prove it first. It sets the benchmark that the construction must meet. Consider an arbitrary algorithm with internal randomness independent of p. For an observed history w (a finite sequence of labels) and a proposed output i, let $c_w(i)$ be the probability, over that internal randomness, that the algorithm stops after precisely this history and returns i. Because the sampler does not know p, this probability is the same for every input distribution. The contribution of this history to the probability of returning i is

$$c_w(i) \prod_{t=1}^{|w|} p_{w_t}, \qquad c_w(i) \ge 0.$$

Suppose a possible stopped history contained only $r < k$ copies of its returned label i. Give i probability ε and assign the remaining mass $1-\varepsilon$ according to a fixed distribution h that gives positive mass to the other labels in the history. This history alone contributes

$$c\varepsilon^r(1-\varepsilon)^{|w|-r}, \qquad c = c_w(i) \prod_{t:\, w_t \ne i} h_{w_t} > 0,$$

whereas exactness requires total output probability

$$q_k(i) = \frac{\varepsilon^k}{\varepsilon^k + (1-\varepsilon)^k \sum_j h_j^k} = O(\varepsilon^k).$$

As $\varepsilon \downarrow 0$, these are incompatible, and other histories cannot cancel a positive contribution. Countability lets us extract such a history whenever, for some p, the returned label has fewer than k copies with positive probability. Hence at least k copies of the returned label must have been observed. Consequently, $N_{\mathcal{B}} \ge J_k$ almost surely on each input (Appendix A.8).

The argument uses universality, since the same algorithm must remain exact as one label's mass varies. It still applies when a finite set of at least two possible labels is known in advance, provided the sampler is exact for every probability vector on it. Because $N_{\mathcal{B}} \ge J_k$ holds almost surely on the same sample stream, the bound also applies to every quantile of the cost, not only its mean. **A collision is therefore necessary, but its winner need not have law q_k.**

4 Turning collisions into exact samples

A first collision supplies both a winner and a finite set of other observed labels. The construction retains this set rather than discarding the information it contains. We treat labels as the states of a Markov chain with stationary law q_k. Each collision block supplies one random step (Section 4.2), and coupling from the past composes steps until the starting state no longer matters (Section 4.3).

[1] On small supports this factor can outweigh the gain. On a fair coin, for example, our sampler costs 5, $33/2$, and $93/2$ samples for $k = 2, 3, 4$, whereas early-abort rejection costs 4, 10, and 22 (Appendix B.1).

4.1 THE FIRST COLLISION IS BIASED

The winner's law already differs from q_k on two labels. Let labels a and b have probabilities u and $1-u$, and consider squaring. Label a is the first to repeat in exactly three ways: (a,a), (a,b,a), or (b,a,a). Its probability of winning is therefore

$$u^2 + 2u^2(1-u) = u^2(3-2u),$$

whereas q_2 gives it probability $u^2/(u^2+(1-u)^2)$. These functions agree at $u = 1/2$ but differ in general. A correct construction must account for the labels encountered before the winner, even though those labels are not returned immediately.

4.2 A COLLISION MAP PRESERVES THE POWERED DISTRIBUTION

Call the samples through J_k, including the final one, a *collision block*. Draw one such block. Its winner y is the label whose count first reaches k. Let $\mathcal{S}$ be all labels seen in the block, and let $A = \mathcal{S} \setminus \{y\}$ be its nonwinning labels. Define the *collision map*

$$F(x) = \begin{cases} x, & x \in A, \\ y, & x \notin A. \end{cases}$$

The map leaves each nonwinning seen label in place, whether seen once or up to $k-1$ times, and sends the winner and every unseen label to the winner. Its image is the finite set $\mathcal{S}$. For example, at $k = 2$ the block (a,b,b) fixes a and sends every other label to b.

Although the winner alone is biased, the map preserves q_k: if $x \sim q_k$ independently of the block, then $F(x) \sim q_k$. For fixed labels x, y, write $P(x,y) = \mathbb{P}(F(x) = y)$ for the transition kernel of a single random map. For distinct positive-mass x, y, the transition $x \to y$ occurs precisely when y wins the collision block and x is absent. For example, at $k = 2$ and for distinct x, y, z, the history (z,y,y) contributes $p_z p_y^2$ to $P(x,y)$. Renaming y as x gives (z,x,x), contributing $p_z p_x^2$ to $P(y,x)$. For general k, because x is absent from every history realizing $x \to y$, renaming y as x pairs these histories one-to-one with those realizing $y \to x$ and replaces the k winner factors p_y^k by p_x^k. Summing over histories gives

$$P(x,y) = p_y^k R_k(x,y), \qquad R_k(x,y) = R_k(y,x). \tag{1}$$

Here $R_k(x,y)$ is the common sum of the remaining factors. Consequently,

$$q_k(x)P(x,y) = \frac{p_x^k p_y^k}{S_k} R_k(x,y) = q_k(y)P(y,x).$$

Thus the kernel is reversible with stationary distribution q_k (Lemma 6). The proof in Appendix A.2 writes R_k as a nonnegative countable sum and needs no finite support. The sampler never computes these probabilities.

4.3 COUPLING FROM THE PAST WITH A FINITE REPRESENTATION

A single random map sends an independent q_k-distributed input to another sample from q_k, but drawing that input is our goal. CFTP removes the need for it by applying the same random maps to every possible starting label (Propp & Wilson, 1996; 1998). Maps from independent collision blocks are drawn one at a time as $F_1, F_2, \ldots$, and each newly drawn map is applied first, as if it acted further in the past. After m maps the composition is $G_m = F_1 \circ \cdots \circ F_m$, and *coalescence* means that G_m is a constant function.

For example, reuse the block (a,b,b) above for F_1, and let F_2 come from (b,c,c), with a, b, c distinct. The new map F_2 fixes b and sends every other label to c. Since F_1 sends both b and c to b, applying F_2 first gives $G_2 = F_1 \circ F_2 \equiv b$.

For each fixed m, the random composition G_m preserves q_k in distribution. Once G_m is constant, every later $G_{m'} = G_m \circ F_{m+1} \circ \cdots \circ F_{m'}$ keeps that value, which makes the stopped output exact (Section 5.3). A forward composition $F_m \circ \cdots \circ F_1$ can change its constant value when the next map is applied last, so stopping it the first time it becomes constant can bias the output.

On an infinite universe the maps cannot be applied label by label, but they need not be: G_m is determined by a default output d and a finite dictionary D of exceptions to that default. For a new map (y, A), the composition $G_m \circ F$ equals $G_m(x)$ on A and $G_m(y)$ outside A. Its representation is therefore

$$d' = G_m(y),$$
$$D' = \{x \mapsto G_m(x) : x \in A,\ G_m(x) \neq d'\}$$

with all evaluations using the old dictionary. In the pseudocode, $\mathrm{lookup}(D, x, d)$ returns $D[x]$ if present and d otherwise. The new default is the output at the sampled winner y, so an empty D' certifies that the composition is constant on all of $\mathcal{X}$, not only on the positive support. Every exception key has been observed, and **coalescence is a finite dictionary test**. Because the default can change, the update must evaluate every new nonwinning label, not only the old exception keys. Appendix A.3 gives an example and proves the representation by induction.

Algorithm 1 Collision-map CFTP

Require: IID oracle draw(), integer $k \geq 2$
function CollisionMap(k)
  $M \leftarrow \varnothing$ ▷ label counts
  repeat
    $y \leftarrow \mathrm{draw}()$
    $M[y] \leftarrow \mathrm{lookup}(M, y, 0) + 1$
  until $M[y] = k$
  return $(y, \mathrm{keys}(M) \setminus \{y\})$
end function
$(y, A) \leftarrow$ CollisionMap(k)
$d \leftarrow y$; $D \leftarrow \{x \mapsto x : x \in A\}$
while $D \neq \varnothing$ **do**
  $(y, A) \leftarrow$ CollisionMap(k)
  $d' \leftarrow \mathrm{lookup}(D, y, d)$
  $D' \leftarrow \varnothing$
  for all $x \in A$ **do**
    $v \leftarrow \mathrm{lookup}(D, x, d)$
    if $v \neq d'$ **then**
      $D'[x] \leftarrow v$
    end if
  end for
  $(d, D) \leftarrow (d', D')$
end while
return d

5 Exactness and sample complexity

We now prove that the backward composition stops, that its output is exact, and that its expected cost is at the collision scale. Let T be the number of collision blocks used. Whether a new block is drawn depends only on the earlier blocks, so each drawn block still has mean length μ_k and the expected cost is $\mathbb{E}N = \mu_k \mathbb{E}T$ (Section 5.3). It therefore suffices to bound μ_k by a multiple of $1/\|p\|_k$ (Section 5.1) and $\mathbb{E}T$ by a constant independent of p (Section 5.2). Appendix A supplies the full proofs.

5.1 The length of one collision block

Poissonization makes label counts independent at each fixed time: attach independent rate-one exponential gaps to the sample stream, giving label i a Poisson process of rate p_i. If τ_k is the first time a label has arrived k times and $\phi_k(t) = e^{-t} \sum_{r=0}^{k-1} t^r/r!$ is the probability that a Poisson variable of mean t is below k, then $\mathbb{P}(\tau_k > t) = \prod_i \phi_k(p_i t)$ (Holst, 1995). A comparison of the logarithms of these survival functions gives

$$e^{-(\|p\|_k t)^k/k!} \leq \mathbb{P}(\tau_k > t) \leq \phi_k(\|p\|_k t). \tag{2}$$

The gaps have mean one, so $\mathbb{E}\tau_k = \mathbb{E}J_k$, and integrating the two bounds yields the collision-time estimates in Theorem 1 (Lemma 5). Poissonization is only an analysis device, and the algorithm draws ordinary samples.

5.2 Coalescence independent of the support size

Invariance alone does not make the maps coalesce, and a bound that tracks every starting label would be useless on an infinite support. Instead we first reduce to a finite image, the set of possible outputs, whose expected size is uniformly bounded.

For analysis, it is convenient to apply maps in forward order, $F_m \circ \cdots \circ F_1$: each new map then acts directly on the current image. Because the maps are IID, this composition has the same law as G_m for each fixed m. Write

$$I_m = \mathrm{im}(F_m \circ \cdots \circ F_1), \qquad K_m = |I_m|.$$

Let $\mathcal{S}_j$ be the labels seen in block j, y_j its winner, and $A_j = \mathcal{S}_j \setminus \{y_j\}$. A map sends every label outside its nonwinning set to its winner, so

$$F_j(I) \subseteq \{y_j\} \cup (I \cap A_j).$$

After k maps, a value that is not one of the later winners must therefore have been seen in every block, and independence makes this event small enough to sum over an arbitrary countable support.

Lemma 3 (Bounded expected image size)**.** *After k independent maps,*

$$\mathbb{E}(K_k - 1) \le k - 2 + \mu_k^k \sum_i p_i^k \le k - 2 + k^k. \tag{3}$$

Proof. A label i appears in a block with probability at most $p_i \mu_k$: its probability of appearing is at most its expected number of appearances, and Wald's identity gives that expectation. Moreover,

$$I_k \subseteq \{y_2, \ldots, y_k\} \cup (\mathcal{S}_1 \cap A_2 \cap \cdots \cap A_k).$$

The expected size of the intersection is at most $\sum_i (p_i \mu_k)^k$, by block independence. Subtracting one from the image size and applying $\|p\|_k \mu_k \le k$ proves the claim. □

The exponent k is essential in this step: *intersecting k independent blocks produces the same power sum that controls their expected lengths*. With fewer blocks, the analogous sum $\mu_k^j \sum_i p_i^j$, $j < k$, can grow with the support, as it does on uniform inputs.

A bounded image can still contain several labels. We therefore need each fresh map to merge any two distinct labels with probability bounded below independently of p. There is a number $\delta_k > 0$, depending only on k, such that

$$\mathbb{P}(F(x) = F(z)) \ge \delta_k \qquad (x \ne z). \tag{4}$$

The bound is uniform in p for the following reason. If $p_x + p_z$ is small relative to $\|p\|_k$, the other labels carry enough kth-power mass to produce a collision while x, z remain unseen, and both labels map to that outside winner. Otherwise, one of the two labels has mass comparable to $\|p\|_k$ and, with probability bounded below in terms of k, wins while the other remains unseen (Appendix A.4).

To turn the pair estimate into an image-size bound, fix any $a \in I_m$. After one more map, every value other than $F(a)$ comes from some $x \in I_m \setminus \{a\}$ that did not merge with a, so $\mathbb{E}[K_{m+1} - 1 \mid I_m] \le (1 - \delta_k)(K_m - 1)$. Together with Lemma 3, this proves

$$\mathbb{P}(K_{k+\ell} > 1) \le (k - 2 + k^k)(1 - \delta_k)^\ell. \tag{5}$$

By the same reversal, the backward composition G_m has the same image-size distribution, and once it is constant, composing it with a further map on the right, $G_m \circ F_{m+1}$, preserves constancy. Hence the backward stopping time T satisfies the same tail bound and

$$\mathbb{E}T \le k + \frac{k - 2 + k^k}{\delta_k}.$$

This gives a finite constant C_k independent of the support. The sharper square-case calculation gives $C_2 = 691/64$ (Appendix A.7).

5.3 An exact stopped output and its expected cost

Having proved termination, we identify the stopped output distribution. Let $\xi \sim q_k$ be independent of all maps, purely for analysis. For every fixed m, invariance implies $G_m(\xi) \sim q_k$. On $\{T \le m\}$, the map G_m is constant and its value is the algorithm's output Z. The coupling inequality therefore gives

$$\mathrm{TV}(\mathcal{L}(Z), q_k) \le \mathbb{P}(T > m) \longrightarrow 0.$$

Thus Z has law exactly q_k. If L_j is the length of the jth collision block, then $\{T \geq j\}$ depends only on the previous blocks, so Wald's identity gives

$$\mathbb{E}N = \sum_{j\geq 1} \mathbb{E}[L_j \mathbf{1}\{T \geq j\}] = \mu_k \sum_{j\geq 1} \mathbb{P}(T \geq j) = \mu_k \mathbb{E}T. \tag{6}$$

The expected number of blocks $\mathbb{E}T$ is therefore exactly the price of exactness relative to the first collision. Counting occurrences within each block and scanning its nonwinning labels gives the dictionary-operation and storage bounds in Theorem 1.

6 Uniform-input asymptotics

The general bounds control the cost factor $\mathbb{E}T$ but do not identify its value. We now determine its limit for fixed k on n equally likely labels as $n \to \infty$, writing T_n, N_n for the block and sample counts. The first-collision winner is already uniform by symmetry. The cost factor $\mathbb{E}T_n$ (Equation (6)) measures *what our sampler pays for not knowing that the input is uniform*.

Each block contains order $n^{1-1/k}$ samples, of which a vanishing fraction are repeats as $n \to \infty$. Its seen and nonwinning sets therefore have the same size as the block to leading order in probability. A label shared by $\mathcal{S}_1, A_2, \ldots, A_m$ remains a possible output: it is in the first image and is fixed by all later maps. Conditional on the set sizes $d_1, \ldots, d_m$, block independence and label symmetry make the sets independent uniform subsets of the n labels. A given label belongs to all of them with probability $\prod_j (d_j/n)$, so the conditional mean intersection size is

$$n \prod_{j=1}^{m} \frac{d_j}{n} = n^{1-m/k} \prod_{j=1}^{m} \frac{d_j}{n^{1-1/k}},$$

which has scale $n^{1-m/k}$. This identifies $m = k$ as the critical step. A variance bound shows that many shared labels survive when $m < k$, preventing coalescence.

To determine the coalescence probability, scale the block lengths by $n^{1-1/k}$. They converge to independent W_j with $\mathbb{P}(W_j > x) = e^{-x^k/k!}$; the corresponding set sizes have the same limits. For set sizes near $n^{1-1/k} w_j$, the shared-label count at $m = k$ converges to a Poisson variable of mean $\prod_{j=1}^{k} w_j$. Averaging its zero probability over the independent limits gives $r_k = \mathbb{E}\exp(-\prod_{j=1}^{k} W_j)$. With probability tending to one, the image after k forward maps consists of these shared labels plus the last winner, which lies outside their intersection. Hence the probability of coalescing by step k tends to r_k. Otherwise one more map suffices with probability tending to one: the expected image size is bounded, and a fresh map leaves a pair distinct only if it sees a nonwinning label from that pair, an event of probability $O(n^{-1/k})$. Thus T_n concentrates on k and $k+1$. Together with the tail bound in Equation (5) and the first-collision mean asymptotic to $m_k n^{1-1/k}$, where $m_k = (k!)^{1/k}\Gamma(1+1/k)$, this gives

$$\mathbb{E}T_n \to k + 1 - r_k, \qquad \mathbb{E}N_n \sim (k + 1 - r_k) m_k n^{1-1/k}.$$

Appendix B.2 gives the full proof. For squaring, the coefficient has a closed form.

Proposition 4 (Uniform-input square cost)**.** *For $k = 2$ and a uniform input on n labels,*

$$\mathbb{P}(T_n = 2) \to \frac{1}{3}, \qquad \mathbb{P}(T_n = 3) \to \frac{2}{3}, \qquad \mathbb{E}T_n \to \frac{8}{3}, \qquad \frac{\mathbb{E}N_n}{\sqrt{n}} \to \frac{8}{3}\sqrt{\frac{\pi}{2}}.$$

For $k = 2$, each W_j has density $xe^{-x^2/2}$ on $x \geq 0$. Averaging the zero-intersection probability over two such lengths gives

$$r_2 = \int_0^\infty \int_0^\infty xy\, e^{-(x+y)^2/2}\, dx\, dy = \frac{1}{3}.$$

The limiting cost factor is therefore $3 - r_2 = 8/3$.

Writing $\mathbb{E}N_{\text{rej}}$ for the expected cost of early-abort rejection on this input, Appendix C.2 gives

$$\mathbb{E}N_{\text{rej}} = n^{k-1}\left(2 + \sum_{j=2}^{k-1} n^{1-j}\right), \qquad \frac{\mathbb{E}N_{\text{rej}}}{\mathbb{E}N_n} \sim \frac{2\, n^{k-2+1/k}}{(k+1-r_k)m_k}.$$

Appendix B.3 develops finite-support recursions and growing-power behavior.

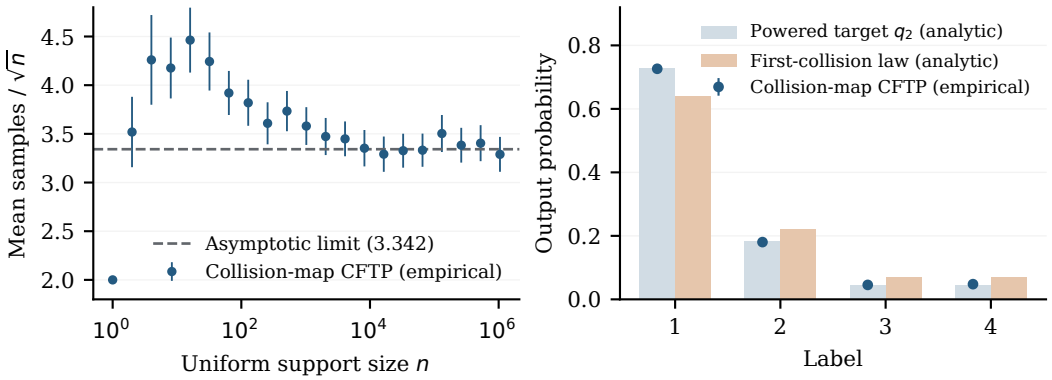


Figure 1: **Left:** Mean sample cost of Collision-map CFTP on uniform inputs (normalized by $\sqrt{n}$), approaching the predicted coefficient $(8/3)\sqrt{\pi/2}$. **Right:** For $p = (1/2, 1/4, 1/8, 1/8)$, bars show the powered target $q_2 = (16, 4, 1, 1)/22$ and the analytically computed output law of the first-collision shortcut, which differs from the target. Points show empirical frequencies from Collision-map CFTP, whose output law is exactly q_2. Error bars indicate two standard errors (in the right panel, they are smaller than the point markers).

6.1 Numerical illustration

At $k = 2$, Figure 1 compares finite-size costs with their asymptotic prediction and empirical output frequencies with the target law. For each uniform support $n = 2^j$, $j = 0, \ldots, 20$, we generate 256 independent outputs. At $n = 2^{20}$ the mean cost is about 3,400 samples (predicted 3,422), whereas early-abort rejection needs $2n$, about two million, in expectation. We also generate 20,000 independent outputs from $p = (1/2, 1/4, 1/8, 1/8)$. For this four-label input, every dictionary update and stopping decision is checked against explicit composition of the sampled maps. Both source oracles use exact uniform random integers.

7 Discussion and conclusion

Exactness is not what makes sample-only powering expensive. Every universal exact sampler must wait until its output label has appeared k times, the biased shortcut of returning the first-collision winner pays exactly that waiting time, and Collision-map CFTP pays it up to a factor depending only on k. The cost is set by how often labels repeat, $1/\|p\|_k$ on each input, not by how rarely k fresh samples agree.

For answer-level sharpening this gives a cost model. Provided each generation is mapped to a label by a fixed extraction rule, the sample cost depends on the distribution of final answers, not on the number of generation paths that produce them. Exact transformations also compose: powering independent q_k outputs by an integer ℓ gives $q_{k\ell}$. Exact samples can serve as a reference for approximate sharpening methods, and potentially as sharpened targets when a teacher model exposes only its outputs.

The guarantees concern one exact output from IID samples at a fixed integer power. For this construction the constant grows at least exponentially with k, while a gambler's-ruin walk that knows the two labels needs only k^2 samples in expectation on a fair coin (Appendix B.1). The optimal dependence on the power is open. Noninteger inverse temperatures could combine an integer stage with a Bernoulli-factory stage (Mendo, 2019) for the fractional part, as in CALM (Shao et al., 2025). Open questions include the collision scale for fractional powers, sharing samples across outputs as in batch settings (Aggarwal et al., 2015), and the cost of approximate sampling.

AI use statement

We used generative AI tools for assisting in the writing of proofs and proposing or refining hypotheses. We have not used generative AI tools for providing critical ingredients for proving mathematical claims, generating synthetic data sets, formulating mathematical claims, helping develop theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, or interpreting results. Cleaning

and reformatting datasets and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-revised text, the LLM-assisted Lean code, and the correspondence between the Lean formalization and the proofs in the manuscript. We take responsibility for the final content of this work.

Ethics statement

This work studies theoretical sampling algorithms and uses synthetic distributions for numerical evaluation. No human participants or personal data are involved.

Reproducibility statement

The sampler, full proofs, and numerical setup are provided in Algorithm 1, Appendices A and B, and Section 6.1. Lean 4 code, numerical code, and supporting data will be released upon publication and can be shared privately with reviewers upon request during the discussion period.

References

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating Rényi entropy of discrete distributions. *IEEE Transactions on Information Theory*, 63(1):38–56, 2017. doi: 10.1109/TIT.2016.2620435.

Divesh Aggarwal, Daniel Dadush, Oded Regev, and Noah Stephens-Davidowitz. Solving the shortest vector problem in 2^n time using discrete gaussian sampling: Extended abstract. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp. 733–742. Association for Computing Machinery, 2015. doi: 10.1145/2746539.2746606. Full version: arXiv:1412.7994.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999. doi: 10.1006/jcss.1997.1545. URL https://web.math.princeton.edu/~nalon/PDFS/amsz4.pdf.

Aleksei Arzhantsev, Otmane Sakhi, and Nicolas Chopin. Self-consistency via marginal sharpening, 2026. URL https://arxiv.org/abs/2605.28142v1. Version 1, May 27, 2026.

Seyedarmin Azizi, Erfan Baghaei Potraghloo, Minoo Ahmadi, Souvik Kundu, and Massoud Pedram. Power-SMC: Low-Latency Sequence-Level Power Sampling for Training-Free LLM Reasoning, 2026. URL https://arxiv.org/abs/2602.10273.

Michael Camarri and Jim Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electronic Journal of Probability*, 5:1–18, 2000. doi: 10.1214/EJP.v5-58. Paper 2.

Fan Chen, Sinho Chewi, Constantinos Daskalakis, and Alexander Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions, 2026. URL https://arxiv.org/abs/2602.01338v2. Version 2.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal Self-Consistency for Large Language Model Generation, 2023. URL https://arxiv.org/abs/2311.17311.

Michael Crouch, Andrew McGregor, Gregory Valiant, and David P. Woodruff. Stochastic streams: Sample complexity vs. space complexity. In *24th Annual European Symposium on Algorithms (ESA 2016)*, volume 57 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 32:1–32:15. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2016. doi: 10.4230/LIPIcs.ESA.2016.32.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 Theorem Prover and Programming Language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pp. 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37. URL https://lean-lang.org/papers/lean4.pdf.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-Dickstein, Kevin Murphy, and Charles Sutton. Language Model Cascades, 2022. URL https://arxiv.org/abs/2207.10342.

Arnaud Doucet, Simon J. Godsill, and Christian P. Robert. Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12(1):77–84, 2002. doi: 10.1023/A:1013172322619.

Shaddin Dughmi, Jason D. Hartline, Robert Kleinberg, and Rad Niazadeh. Bernoulli factories and black-box reductions in mechanism design. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 158–169. Association for Computing Machinery, 2017. doi: 10.1145/3055399.3055492.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

James Allen Fill. An interruptible algorithm for perfect sampling via Markov chains. *The Annals of Applied Probability*, 8 (1):131–162, 1998. doi: 10.1214/aoap/1027961037. URL https://www.ams.jhu.edu/~fill/papers/algo1.pdf.

James Allen Fill, Motoya Machida, Duncan J. Murdoch, and Jeffrey S. Rosenthal. Extension of Fill's perfect rejection sampling algorithm to general chains. *Random Structures & Algorithms*, 17(3–4):290–316, 2000. doi: 10.1002/1098-2418(200010/12)17:3/4<290::AID-RSA6>3.0.CO;2-Q.

Philippe Flajolet, Danièle Gardy, and Loÿs Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229, 1992. doi: 10.1016/0166-218X(92)90177-C. URL https://algo.inria.fr/flajolet/Publications/FlGaTh92.pdf.

S. G. Foss and R. L. Tweedie. Perfect simulation and backward coupling. *Communications in Statistics. Stochastic Models*, 14(1–2):187–203, 1998. doi: 10.1080/15326349808807466.

Lars Holst. The general birthday problem. *Random Structures & Algorithms*, 6(2–3):201–208, 1995. doi: 10.1002/rsa.3240060207.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Mark Huber. Perfect sampling using bounding chains. *The Annals of Applied Probability*, 14(2):734–753, 2004. doi: 10.1214/105051604000000080.

Mark Huber. Nearly optimal Bernoulli factories for linear functions. *Combinatorics, Probability and Computing*, 25(4): 577–591, 2016. doi: 10.1017/S0963548315000371.

Mark Huber. Optimal linear Bernoulli factories for small mean problems. *Methodology and Computing in Applied Probability*, 19(2):631–645, 2017. doi: 10.1007/s11009-016-9518-3.

Gautam Iyer. Markov chains: Introduction. Carnegie Mellon University, 21-326 course notes, 2025. URL https://www.math.cmu.edu/~gautam/c/2025-326/notes/markov1.html. Examples 15 and 20: gambler's ruin probabilities and absorption times. Accessed September 19, 2026.

Rajesh Jayaram and David P. Woodruff. Perfect L_p sampling in a data stream. *SIAM Journal on Computing*, 50(2): 382–439, 2021. doi: 10.1137/18M1229912.

Rajesh Jayaram, David P. Woodruff, and Samson Zhou. Truly perfect samplers for data streams and sliding windows. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 29–40. Association for Computing Machinery, 2022. doi: 10.1145/3517804.3524139.

Xiaotong Ji, Rasul Tutunov, Matthieu Zimmer, and Haitham Bou Ammar. Scalable Power Sampling: Unlocking Efficient, Training-Free Reasoning for LLMs via Distribution Sharpening, 2026. URL https://arxiv.org/abs/2601.21590.

Aayush Karan and Yilun Du. Reasoning with Sampling: Your Base Model is Smarter Than You Think. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/01498f17142547dadb60999c39bd0a09-Abstract-Conference.html.

M. S. Keane and George L. O'Brien. A Bernoulli factory. *ACM Transactions on Modeling and Computer Simulation*, 4(2): 213–219, 1994. doi: 10.1145/175007.175019.

Wilfrid S. Kendall and Jesper Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, 32(3):844–865, 2000. doi: 10.1239/aap/1013540247.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Krzysztof Łatuszyński, Ioannis Kosmidis, Omiros Papaspiliopoulos, and Gareth O. Roberts. Simulating events of unknown probabilities via reverse time martingales. *Random Structures & Algorithms*, 38(4):441–452, 2011. doi: 10.1002/rsa.20333.

Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K. Mansinghka. Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs, 2023. URL https://arxiv.org/abs/2306.03081.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Luis Mendo. An asymptotically optimal Bernoulli factory for certain functions that can be expressed as power series. *Stochastic Processes and their Applications*, 129(11):4366–4384, 2019. doi: 10.1016/j.spa.2018.11.017.

Giulio Morina, Krzysztof Łatuszyński, Piotr Nayar, and Alex Wendland. From the Bernoulli factory to a dice enterprise via perfect sampling of Markov chains. *The Annals of Applied Probability*, 32(1):327–359, 2022. doi: 10.1214/21-AAP1679.

D. J. Murdoch and P. J. Green. Exact sampling from a continuous state space. *Scandinavian Journal of Statistics*, 25(3): 483–502, 1998. doi: 10.1111/1467-9469.00116.

Şerban Nacu and Yuval Peres. Fast simulation of new coins from old. *The Annals of Applied Probability*, 15(1A):93–115, 2005. doi: 10.1214/105051604000000549.

Rad Niazadeh, Renato Paes Leme, and Jon Schneider. Combinatorial Bernoulli factories. *Bernoulli*, 29(2):1246–1274, 2023. doi: 10.3150/22-BEJ1497.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large Language Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/48b383b24230e0e6e649d9c98dae4d8c-Abstract-Conference.html.

Maciej Obremski and Maciej Skorski. Rényi entropy estimation revisited. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*, volume 81 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 20:1–20:15. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.APPROX-RANDOM.2017.20.

Renato Paes Leme and Jon Schneider. Note on finite-automata Bernoulli factories for rational functions, 2026. URL https://arxiv.org/abs/2606.29595v1. Version 1, June 28, 2026.

James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1–2):223–252, 1996. doi: 10.1002/(SICI)1098-2418(199608/09)9:1/2<223::AID-RSA14>3.0.CO;2-O.

James Gary Propp and David Bruce Wilson. How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *Journal of Algorithms*, 27(2):170–217, 1998. doi: 10.1006/jagm.1997.0917. URL https://www2.stat.duke.edu/~scs/Projects/Trees/Theory/ProppWilson1998.pdf.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561. University of California Press, 1961. URL https://static.renyi.hu/renyi_cikkek/1961_on_measures_of_entropy_and_information.pdf.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Sebastian M. Schmon, Arnaud Doucet, and George Deligiannidis. Bernoulli race particle filters. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2350–2358. PMLR, 2019. URL https://proceedings.mlr.press/v89/schmon19a.html.

Chenze Shao, Darren Li, Fandong Meng, and Jie Zhou. Continuous autoregressive language models, 2025. URL https://arxiv.org/abs/2510.27688v1. Version 1, October 31, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

The mathlib Community. The Lean Mathematical Library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pp. 367–381. Association for Computing Machinery, 2020. doi: 10.1145/3372885.3373824. URL https://leanprover-community.github.io/papers/mathlib-paper.pdf.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

David Bruce Wilson. How to couple from the past using a read-once source of randomness. *Random Structures & Algorithms*, 16(1):85–113, 2000. doi: 10.1002/(SICI)1098-2418(200001)16:1<85::AID-RSA6>3.0.CO;2-H.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Grosse. Probabilistic Inference in Language Models via Twisted Sequential Monte Carlo. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60704–60748. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24c.html.

Felix Zhou, Anay Mehrotra, and Quanquan C. Liu. Reasoning with Sampling: Cutting at Decision Points, 2026. URL https://arxiv.org/abs/2605.30327.

APPENDIX CONTENTS

A Complete proofs of the main results . . . 13
A.1 First-collision cost on countable support . . . 13
A.2 Stationarity of one map . . . 14
A.3 Finite representation of backward composition . . . 14
A.4 A uniform pair-coalescence bound for each fixed power . . . 15
A.5 Bounded image size and coalescence . . . 15
A.6 Exact stopped output and total work . . . 16
A.7 The square-case constant . . . 17
A.8 Universal exactness forces k observations of the output . . . 17
A.9 The uniform-input square limit . . . 18
B Theoretical extensions and dependence on the power . . . 18
B.1 An unavoidable large-power limitation of this construction . . . 18
B.2 Uniform-input asymptotics for fixed integer powers . . . 19
B.3 Finite uniform sources and joint growth . . . 22
C Connections to exact distribution transformations . . . 24
C.1 Connections to other exact-sampling frameworks . . . 24
C.2 Rejection based on identical observations . . . 24
C.3 Sampling interfaces and applications . . . 24
D Formal verification in Lean 4 . . . 25

A COMPLETE PROOFS OF THE MAIN RESULTS

This appendix supplies complete proofs of the main guarantees and the square-case limit. Appendix B develops the further asymptotic and finite-support results.

Throughout, $k \geq 2$ is fixed, p is a probability distribution on a countable universe $\mathcal{X}$, $S_k = \sum_i p_i^k$, and $a_k = \|p\|_k = S_k^{1/k} > 0$. We count the final threshold-crossing draw in every block. The actual algorithm uses neither the probabilities nor the Poisson clocks introduced below.

A.1 FIRST-COLLISION COST ON COUNTABLE SUPPORT

Lemma 5 (Two-sided collision-time bound)**.** *The k-fold collision time J_k is almost surely finite, and*

$$\frac{(k!)^{1/k}\Gamma(1+1/k)}{a_k} \leq \mu_k = \mathbb{E}[J_k] \leq \frac{k}{a_k}.$$

Proof. Any fixed label of positive probability eventually appears k times, so its kth appearance bounds J_k from above almost surely. Independently attach IID rate-one exponential interarrival times to the complete label stream. Poisson marking gives independent label processes of rates p_i. Let τ_k be the first time any process reaches k arrivals. Put

$$E_{k-1}(z) = \sum_{r=0}^{k-1} \frac{z^r}{r!}, \qquad \phi_k(z) = e^{-z}E_{k-1}(z), \qquad f_k(z) = -\log \phi_k(z).$$

Finite-subset independence and continuity of probability for decreasing events give $\mathbb{P}(\tau_k > t) = \prod_i \phi_k(p_i t)$. This also covers finite support and zero-mass labels. Direct differentiation and integration yield

$$f_k'(z) = \frac{z^{k-1}}{(k-1)!E_{k-1}(z)},$$
$$\frac{f_k(z)}{z^k} = \frac{1}{(k-1)!}\int_0^1 \frac{u^{k-1}}{E_{k-1}(zu)}\,du.$$

The second expression, continuously extended to $z = 0$, is decreasing. Moreover $0 \leq f_k(z) \leq z^k/k!$. In particular $\sum_i f_k(p_i t) \leq (a_k t)^k/k! < \infty$, so the product has a finite logarithm for finite t. Because $p_i \leq a_k$, monotonicity gives

$$\sum_i f_k(p_i t) \geq f_k(a_k t) \sum_i \frac{p_i^k}{a_k^k} = f_k(a_k t).$$

Hence

$$\exp\left(-\frac{(a_k t)^k}{k!}\right) \le \mathbb{P}(\tau_k > t) \le \phi_k(a_k t).$$

The exponential interarrival times are independent of J_k, which is determined by labels alone. Tonelli therefore gives $\mathbb{E}[\tau_k] = \mathbb{E}[J_k]$, even before finiteness is known. Integrating the lower survival bound gives $(k!)^{1/k}\Gamma(1+1/k)/a_k$. Each of the k summands in $e^{-a_k t}E_{k-1}(a_k t)$ integrates to $1/a_k$, proving the upper bound. □

The survival argument also applies to any subcollection of label processes, with a_k replaced by that subcollection's ℓ_k norm. Their rates need not sum to one. This observation will be used for processes outside a selected pair.

A.2 Stationarity of one map

Lemma 6 (Detailed balance)**.** *Let F be the collision map, and let $P(x,y) = \mathbb{P}(F(x) = y)$. Then $q_k(i) = p_i^k / S_k$ is invariant for P.*

Proof. Work first with distinct positive-mass labels x, y. The event $F(x) = y$ requires that y win the block and that x never appear. Before the final y, there are $k-1$ copies of y and $c_j \in \{0, \dots, k-1\}$ copies of each other label $j \notin \{x, y\}$. Only finitely many c_j are nonzero. Every ordering of this prefix is legal: no count can have reached k. Writing $|c| = \sum_j c_j$, the number of orderings is $(k-1+|c|)!/((k-1)!\prod_j c_j!)$. Thus

$$P(x,y) = p_y^k R_k(x,y),$$
$$R_k(x,y) = \sum_{\substack{c \text{ finitely supported} \\ c_j \in \{0,\dots,k-1\},\ j \notin \{x,y\}}} \frac{(k-1+|c|)!}{(k-1)!\prod_j c_j!} \prod_j p_j^{c_j}.$$

Products here run only over the positive entries of c. The coefficient is symmetric in x, y. The enumeration is a countable sum of disjoint finite-word events, so it is valid on countable support by nonnegative summation. It is finite when $p_y > 0$ since $P(x,y) \le 1$. Consequently

$$q_k(x)P(x,y) = \frac{p_x^k p_y^k}{S_k} R_k(x,y) = q_k(y)P(y,x).$$

Diagonal balance is automatic. The image of every sampled map consists only of sampled, positive-mass labels, so transitions cannot output a zero-mass label. Summing the nonnegative balance identities proves $q_k P = q_k$ on the entire universe. □

The lemma concerns the transition produced by applying a random map to a state. It does not identify the distribution of the block's winning label.

A.3 Finite representation of backward composition

Lemma 7 (Exact default-and-exception update)**.** *Algorithm 1 represents $G_m = F_1 \circ \cdots \circ F_m$ exactly. Its dictionary is empty if and only if G_m is constant on the positive support, or equivalently on all of $\mathcal{X}$. Once the dictionary is empty, composing G_m with further (older) maps leaves its constant value unchanged.*

Proof. For the first map, $d = y_1$ and $D[x] = x$ for $x \in A_1$ represent F_1, with no stored output equal to d. Suppose the old representation is exact. For a fresh map F with winner y and nonwinning set A, every $x \notin A$ satisfies $(G_m \circ F)(x) = G_m(y) = d'$. Every $x \in A$ instead has output $G_m(x)$, giving precisely the dictionary update displayed in Section 4.3. In particular, it must inspect every $x \in A$, not only old exception keys.

The default d' is attained at y, since $y \notin A$ and $F(y) = y$. The winner is sampled and has positive probability. Every new exception key belongs to A, so it too is a sampled, positive-mass label. A nonempty dictionary therefore certifies two distinct outputs attained on the positive support, while

an empty dictionary represents a constant function everywhere. Finally, if G_m is constant, $G_m \circ F$ is the same constant for every older map F. □

For example, an old default a with the sole exception $b \mapsto b$, followed by a fresh winner b and nonwinning set $\{c\}$ for distinct a, b, c, produces new default b and the necessary exception $c \mapsto a$. The key c was absent from the old dictionary. Restricting the scan to old exceptions would incorrectly declare coalescence.

A.4 A UNIFORM PAIR-COALESCENCE BOUND FOR EACH FIXED POWER

Define

$$\psi_k(v) = 1 - e^{-v} E_{k-1}(v), \qquad d_k = (1 - 2^{-k})^{1/k},$$

and the positive constants

$$\delta_k = \min\left\{ \psi_k(k) \exp\left(-\frac{k}{2d_k}\right),\ \psi_k(1/4) \exp\left(-1 - \frac{1}{k!}\right) \right\}, \qquad \rho_k = 1 - \delta_k < 1.$$

These are coupling constants, not accuracy parameters. In particular, $\psi_k(1/4)$ makes this general bound very poor as k grows.

Lemma 8 (Pair contraction)**.** *For any distinct labels x, z and any p, $\mathbb{P}(F(x) \neq F(z)) \leq \rho_k$.*

Proof. Put $a = a_k$, $s = p_x + p_z$, and $a_{\text{out}}^k = \sum_{i \notin \{x,z\}} p_i^k$. Use the independent Poisson label processes from Lemma 5.

If $s \leq a/2$, then $a_{\text{out}}^k \geq a^k - s^k \geq a^k(1 - 2^{-k})$, so $a_{\text{out}} \geq a d_k > 0$. At $t = k/a_{\text{out}}$, the outside processes have reached a kth occurrence with probability at least $\psi_k(k)$. Independently, neither x nor z has arrived by t with probability $e^{-st} \geq e^{-k/(2d_k)}$. On the intersection, the global winner is an outside label and both x, z are unseen, hence both are sent to that winner.

If $s > a/2$, let $u = \max(p_x, p_z) > a/4$ and $v = \min(p_x, p_z) \leq a$, retaining their associated label names. At $t = 1/a$, require at least k arrivals of the u label, no arrival of the v label, and fewer than k arrivals of every outside label. The independent group probabilities are at least $\psi_k(1/4)$, e^{-1}, and

$$\exp\left(-\frac{t^k a_{\text{out}}^k}{k!}\right) \geq e^{-1/k!},$$

respectively. For empty outside support the third probability is one. The u label wins before t, the other pair label is unseen, and both map to the u label. The minimum of these two success bounds proves the claim. Zero-mass pair labels cause no difficulty in the same cases. Poisson arrival ties have probability zero, and all events used concern finite times. □

A.5 BOUNDED IMAGE SIZE AND COALESCENCE

Let $F_1, F_2, \ldots$ be independent maps applied in forward order, and set

$$I_m = \text{im}(F_m \circ \cdots \circ F_1), \qquad K_m = |I_m|.$$

Each I_m is finite and nonempty because $I_1 = \mathcal{S}_1$.

Lemma 9 (Independent blocks remove support-size dependence)**.** *With $b_k = k - 2 + k^k$, for every integer $\ell \geq 0$,*

$$\mathbb{E}[K_k - 1] \leq b_k, \qquad \mathbb{E}[K_{k+\ell} - 1] \leq b_k \rho_k^\ell.$$

Proof. Since $F_m(I) \subseteq \{y_m\} \cup (I \cap A_m)$, induction gives

$$I_m \subseteq \{y_2, \ldots, y_m\} \cup (\mathcal{S}_1 \cap A_2 \cap \cdots \cap A_m), \qquad m \geq 2.$$

For any label i, a stopped-sum calculation gives

$$\mathbb{P}(i \in \mathcal{S}) \leq \mathbb{E}\left[\sum_{r=1}^{J_k} \mathbf{1}\{X_r = i\}\right] = \sum_{r \geq 1} p_i \mathbb{P}(J_k \geq r) = p_i \mu_k.$$

The event $\{J_k \geq r\}$ is determined before draw r, so including the final draw is consistent with the equality. Since $A \subseteq \mathcal{S}$, block independence and Tonelli imply

$$\mathbb{E}|\mathcal{S}_1 \cap A_2 \cap \cdots \cap A_k| \leq \sum_i (p_i \mu_k)^k = (a_k \mu_k)^k \leq k^k.$$

There are at most $k-1$ displayed winner labels, yielding the first assertion.

For a finite nonempty set I and any chosen anchor $x \in I$,

$$|F(I)| - 1 \leq \sum_{z \in I \setminus \{x\}} \mathbf{1}\{F(z) \neq F(x)\}.$$

Conditioning on I_m and using a fresh independent map with Lemma 8 gives $\mathbb{E}[K_{m+1} - 1 \mid I_m] \leq \rho_k(K_m - 1)$. Iteration proves the second assertion. Only finite image sets are used in this contraction, so there is no union bound over the unknown countable support. □

Lemma 10 (Finite backward coalescence)**.** *For the stopping map count T of Algorithm 1,*

$$\mathbb{P}(T > k + \ell) \leq \min\{1, b_k \rho_k^\ell\},$$

$$\mathbb{E}[T] \leq k + \frac{b_k}{\delta_k} < \infty.$$

Proof. For every deterministic m, reversing m IID maps preserves their joint law. Thus $G_m = F_1 \circ \cdots \circ F_m$ and the forward composition have the same image-cardinality distribution. By Lemma 7, backward coalescence is absorbing, so $\{T > m\}$ is exactly the event that G_m is not constant. Markov's inequality and Lemma 9 give the tail bound. Summing it yields

$$\begin{aligned}\mathbb{E}[T] &= \sum_{m \geq 0} \mathbb{P}(T > m) \\ &\leq k + \sum_{\ell \geq 0} \min\{1, b_k \rho_k^\ell\} \leq k + \frac{b_k}{1 - \rho_k}.\end{aligned}$$

In particular T is finite almost surely. □

A.6 Exact stopped output and total work

Proof of Theorem 1, except the sharper square constant. Let Z be the algorithm's almost surely finite output. Introduce a conceptual state $\xi \sim q_k$ independent of the entire map sequence. For every fixed m, Lemma 6 implies $G_m(\xi) \sim q_k$. On $\{T \leq m\}$, the composition is constant and $G_m(\xi) = Z$. Consequently

$$\mathrm{TV}(\mathcal{L}(Z), q_k) \leq \mathbb{P}(G_m(\xi) \neq Z) \leq \mathbb{P}(T > m) \longrightarrow 0.$$

This is a fixed-time coupling argument, not an optional-stopping claim about a stationary chain. The conceptual state is not sampled by the algorithm.

Successive collision blocks are IID, jointly in their lengths and maps, because each block ends at an almost surely finite stopping index, after which the IID stream starts afresh. Let $J_{k,r}$ be block r's length. The request event $\{T \geq r\}$ depends only on preceding completed blocks and is independent of the fresh $J_{k,r}$. Hence Tonelli gives

$$\mathbb{E}[N] = \sum_{r \geq 1} \mathbb{E}\big[J_{k,r} \mathbf{1}\{T \geq r\}\big] = \mu_k \sum_{r \geq 1} \mathbb{P}(T \geq r) = \mu_k \mathbb{E}[T].$$

A block length and its own map are generally correlated, so the proof does not assume their independence. Nor is it valid to condition on T and assign mean μ_k to each resulting conditioned length. Lemmas 5 and 10 give the asserted bounds with $C_k = k + b_k/\delta_k$, before the square-case refinement below.

Building a block's count dictionary and scanning its complete nonwinning set requires $O(J_{k,r})$ dictionary operations. Maintaining old and new exception dictionaries during the update changes only the constant. The expected operation count is therefore $O(\mathbb{E}[N])$. Expected constant-time hash operations imply the same bound for bookkeeping time under the unit-cost label model. Peak stored entries are bounded by a constant times the total draw count, so expected peak storage is also $O_k(a_k^{-1})$. These are not worst-case hashing guarantees, byte-memory bounds for arbitrary-length labels, or claims about oracle generation latency. □

A.7 THE SQUARE-CASE CONSTANT

Lemma 11 (Nonwinning labels for squaring)**.** *For $k = 2$, every label belongs to A with probability at most $1/e$, and each distinct pair remains separate under a fresh map with probability at most $\rho = 2/e < 3/4$.*

Proof. If $0 < p_i < 1$, let τ_{-i} be the first repeat time among all other label processes. It is finite almost surely and independent of label i's Poisson process. The event $i \in A$ is exactly one i arrival by τ_{-i}. Thus

$$\mathbb{P}(i \in A) = \mathbb{E}\big[p_i \tau_{-i} e^{-p_i \tau_{-i}}\big] \le 1/e.$$

For $p_i = 0$ the probability of membership in A is zero, and for $p_i = 1$ the map is constant, so the bound remains valid. For distinct x, z, their map images differ exactly when at least one belongs to A. A union bound, without assuming independence of these two membership events, gives $2/e$. □

Square refinement in Theorem 1. The two-map image argument gives $\mathbb{E}[K_2 - 1] \le (a_2\mu_2)^2 \le 4$. Use Lemma 11 for subsequent contraction, and retain the probability cap in the tail sum:

$$\begin{aligned}\mathbb{E}[T] &\le 2 + \sum_{\ell \ge 0} \min\{1, 4(2/e)^\ell\} \\ &\le 2 + 5 + \frac{4(3/4)^5}{1 - 3/4} = \frac{691}{64} < 11.\end{aligned}$$

Combining the stopped-sum cost identity with $\mu_2 \le 2/a_2$ proves the claimed square bound. □

The one-label union-bound argument is specific to squaring. For larger k, the probability of between one and $k - 1$ arrivals can exceed $1/2$, so doubling a bound on membership in A need not yield a positive pair-coalescence probability. Lemma 8 uses joint events to avoid that issue.

A.8 UNIVERSAL EXACTNESS FORCES k OBSERVATIONS OF THE OUTPUT

Proof of Theorem 2. Fix one universal algorithm whose internal randomness is independent of p. For a finite word $w = (w_1, \dots, w_n)$ and proposed output i, let $c_w(i)$ be the probability over its internal randomness that, when supplied prefix w, it stops after exactly n oracle requests and returns i. This is a measurable event probability independent of p. No finite-bit or interval assumption about internal randomness is required. Its stopped-transcript probability is

$$c_w(i) \prod_{r=1}^{n} p_{w_r}.$$

This includes $n = 0$ and the empty word. If, for some fixed p, output i has fewer than k occurrences with positive probability, countability of $\mathcal{X}$ and of all finite words gives some i, w with $c_w(i) > 0$, positive word probability, and $r < k$ occurrences of i.

Choose a fixed distribution h on $\mathcal{X} \setminus \{i\}$ assigning positive mass to each other label appearing in w. If no such label appears, choose any other possible label, available because $|\mathcal{X}| \ge 2$. For $0 < \varepsilon < 1$, define

$$p_\varepsilon(i) = \varepsilon, \qquad p_\varepsilon(j) = (1 - \varepsilon)h_j \quad (j \ne i).$$

The chosen transcript has probability $c\varepsilon^r(1 - \varepsilon)^{n-r}$ for a constant $c > 0$. Exactness for this family instead requires

$$\mathbb{P}_{p_\varepsilon}(\text{output} = i) = \frac{\varepsilon^k}{\varepsilon^k + (1 - \varepsilon)^k \sum_j h_j^k} = O(\varepsilon^k).$$

As $\varepsilon \downarrow 0$, the single nonnegative transcript contribution contradicts this requirement when $r < k$. Other transcript probabilities cannot cancel it. Therefore, for each fixed p, the returned label has appeared at least k times almost surely. Along the same IID stream this implies $N_{\mathcal{B}} \ge J_k$, and taking expectations proves the result. □

The comparison allows a known finite alphabet, provided arbitrary probability vectors on it remain in the required family. It excludes distribution-dependent advice and restricted promises such as uniformity. A known singleton universe is excluded, whereas an actual point mass in the unrestricted at-least-two-label universe is included and still forces k observations for a universal exact algorithm. All observations used before the fresh output must be counted. The proof is not an amortized lower bound for multiple outputs and does not extend to approximate outputs or additional likelihood access.

A.9 The uniform-input square limit

The following proves Proposition 4. The functions ϕ_k are defined in Appendix A.1, and K_m is the forward image size from Appendix A.5.

Proof. Write J_n for one block's draw count, τ_n for its Poissonized time, and $\mu_n = \mathbb{E}J_n$. Then $\mathbb{P}(\tau_n > \sqrt{n}\,t) = \phi_2(t/\sqrt{n})^n \to e^{-t^2/2}$, and these probabilities are bounded by the integrable $\phi_2(t)$, so dominated convergence gives $\mu_n/\sqrt{n} \to \sqrt{\pi/2}$. Conditional on the label-determined J_n, the Poissonized time is a sum of J_n independent exponential gaps, so $\mathbb{E}(\tau_n - J_n)^2 = \mu_n = O(\sqrt{n})$. It follows that $J_n/\sqrt{n} \Rightarrow W$ with survival $e^{-t^2/2}$.

In a square block, $|\mathcal{S}| = J_n - 1$ and $|A| = J_n - 2$. Independent blocks give independent uniform subsets conditional on their sizes $d_1 = |\mathcal{S}_1|$ and $d_2 = |A_2|$. For $Z_n = |\mathcal{S}_1 \cap A_2|$,

$$\mathbb{P}(Z_n = 0 \mid d_1, d_2) = \frac{\binom{n-d_1}{d_2}}{\binom{n}{d_2}} = \prod_{j=0}^{d_2-1}\left(1 - \frac{d_1}{n-j}\right) \longrightarrow e^{-x_1 x_2}$$

whenever $d_j/\sqrt{n} \to x_j > 0$. The logarithm converges uniformly on compact sets of (x_1, x_2), and tightness of the two normalized sizes then gives the unconditional limit $\mathbb{E}e^{-W_1 W_2}$ for independent copies W_1, W_2 of W.

The second forward image is $(\mathcal{S}_1 \cap A_2) \cup \{y_2\}$ unless $\mathcal{S}_1 \subseteq A_2$. Choosing an anchor in $\mathcal{S}_1$ bounds this exception's probability by $\alpha_n = \mathbb{E}|A|/n = O(n^{-1/2})$. Since $y_2 \notin A_2$, outside this exception $K_2 = Z_n + 1$, so $\mathbb{P}(K_2 = 1) \to r_2$, where

$$r_2 = \mathbb{E}e^{-W_1 W_2} = \int_0^\infty \int_0^\infty xy e^{-(x+y)^2/2}\,dx\,dy = \frac{1}{6}\int_0^\infty u^3 e^{-u^2/2}\,du = \frac{1}{3}.$$

A fresh map separates a fixed pair with probability at most $\xi_n = 2\alpha_n = O(n^{-1/2})$. For sufficiently large n, the anchored-image argument and $\mathbb{E}[K_2 - 1] \le 4$ give

$$\sum_{m \ge 3} \mathbb{P}(T_n > m) \le \frac{4\xi_n}{1 - \xi_n} \longrightarrow 0.$$

Finally, $\mathbb{P}(T_n = 1) = 1/n$ and $\mathbb{P}(T_n \le 2) = \mathbb{P}(K_2 = 1) \to 1/3$ by fixed-time reversal. The vanishing tail proves the two-point law and $\mathbb{E}T_n \to 8/3$. Equation (6) yields the asserted sample-cost limit. □

B Theoretical extensions and dependence on the power

The guarantees above fix k and allow constants depending on it. We now examine that dependence, first on two labels, then on large uniform supports at fixed k, and finally when k grows with the support size.

B.1 An unavoidable large-power limitation of this construction

Proposition 12 (Two-label boundary)**.** *Suppose the positive support is $\{a, b\}$, with probabilities p and $1 - p$, where $0 < p < 1$. For Algorithm 1,*

$$T \sim \mathrm{Geom}\big(p^k + (1-p)^k\big), \qquad \mathbb{E}N = \frac{\mu_k}{p^k + (1-p)^k},$$

where the geometric distribution has support $\{1, 2, \ldots\}$. *In particular, on the fair two-label law,* $\mathbb{E}T = 2^{k-1}$. *Thus any bound* $\mathbb{E}N \le C_k \mu_k$ *for this construction over all input laws must have* $C_k \ge 2^{k-1}$.

Proof. A block that sees both labels fixes the nonwinner and also maps its winner to itself. Its restriction to the positive support is therefore the identity. A block seeing only one label consists of k identical draws and produces a constant map. A block is constant with probability $p^k + (1-p)^k$, independently of the other blocks. Until the first such block, a composition remains the identity on $\{a, b\}$, and the first constant block makes it constant.

This argument also applies when the ambient universe contains infinitely many zero-mass labels. Every exception key belongs to the positive support, and the stored default is attained at a positive-mass label. Thus the exception dictionary is empty exactly when the composition is constant on the positive support, so zero-mass labels cannot delay or prematurely trigger termination. The map count is geometric as stated, and the stopped-sum identity gives its expected observation cost. □

Since $k \le J_k \le 2k - 1$ on two labels, the fair-law cost is $\Theta(k2^k)$. Each early-abort rejection trial has expected length $\ell_k = 2 + S_2 + \cdots + S_{k-1}$, so early-abort rejection costs ℓ_k / S_k in expectation. The CFTP-to-rejection cost ratio is $\mu_k/\ell_k > 1$ for every nondegenerate two-label law. For $k > 2$, this follows from $\mu_k \ge k > \ell_k$, and for $k = 2$ from $\mu_2 > 2 = \ell_2$. On the fair law, the exact CFTP costs for $k = 2, 3, 4$ are 5, $33/2$, and $93/2$, respectively, versus early-abort costs of 4, 10, and 22. These exact values show that constant-factor optimality need not imply a lower cost than early-abort rejection on every input.

Consequently, very low temperatures, that is, large k, remain costly for this construction even though its dependence on the input is optimal for each fixed k. The exponential dependence is *a limitation of the collision-map construction*, not a lower bound for every exact sampler. For comparison, on a *known two-label alphabet*, the classical gambler's-ruin walk starts at zero, adds $+1$ for a and -1 for b, and returns the label of the first boundary hit at $\pm k$ (Iyer, 2025). If $h(s)$ is the probability of hitting $+k$ first, then

$$h(-k) = 0, \quad h(k) = 1, \quad h(s) = ph(s+1) + (1-p)h(s-1).$$

For $0 < p < 1$ and $p \ne 1/2$, writing $\rho = (1-p)/p$ gives

$$h(0) = \frac{1 - \rho^k}{1 - \rho^{2k}} = \frac{p^k}{p^k + (1-p)^k},$$

and the fair case has $h(0) = 1/2$. At $p = 1/2$, the mean absorption time from s is $k^2 - s^2$, as follows from the first-step cost recurrence and zero boundary values, so the fair walk started at zero uses k^2 observations in expectation. At $p = 0$ or $p = 1$, the walk stops deterministically after k observations and returns the sole positive-mass label. The same walk is thus exact for all probability vectors on this known two-label alphabet. This example still satisfies the output-multiplicity lower bound, since reaching a boundary requires at least k occurrences of the returned label. It is not, however, a single algorithm for an unknown or countably infinite support.

B.2 Uniform-input asymptotics for fixed integer powers

Proposition 13 (Uniform-input asymptotics)**.** *Fix an integer* $k \ge 2$ *and let* p *be uniform on* n *labels. Write* T_n, N_n *for the map and draw counts of Algorithm 1. Let* $W_1, \ldots, W_k$ *be independent with survival* $\mathbb{P}(W_j > x) = e^{-x^k/k!}$ *for* $x \ge 0$*, and define*

$$m_k = (k!)^{1/k}\Gamma(1 + 1/k), \qquad r_k = \mathbb{E}\exp\left(-\prod_{j=1}^{k} W_j\right).$$

As $n \to \infty$ *with* k *fixed,*

$$\mathbb{P}(T_n = k) \to r_k, \qquad \mathbb{P}(T_n = k+1) \to 1 - r_k, \qquad \mathbb{E}T_n \to k + 1 - r_k,$$
$$\frac{\mathbb{E}N_n}{n^{1-1/k}} \to (k + 1 - r_k)m_k.$$

In particular, $r_2 = 1/3$ *and* $\mathbb{E}N_n/\sqrt{n} \to (8/3)\sqrt{\pi/2}$ *for* $k = 2$.

Proof of Proposition 13. Fix $k \geq 2$ and let p be uniform on n labels. Write $b_n = n^{1-1/k}$, J_n for one block's draw count, $\mu_n = \mathbb{E}J_n$, and T_n, N_n for the map and total draw counts. Let $W_1, \ldots, W_k$ be independent with $\mathbb{P}(W_j > x) = e^{-x^k/k!}$, $x \geq 0$, and put

$$m_k = (k!)^{1/k}\Gamma(1 + 1/k), \qquad \Lambda_k = \prod_{j=1}^{k} W_j, \qquad r_k = \mathbb{E}e^{-\Lambda_k}.$$

Block lengths and seen sets. Use the independent exponential gaps and functions ϕ_k, f_k from Lemma 5. The Poissonized stopping time τ_n satisfies

$$\mathbb{P}(\tau_n > b_n x) = e^{-nf_k(xn^{-1/k})} \longrightarrow e^{-x^k/k!}, \qquad \mathbb{P}(\tau_n > b_n x) \leq \phi_k(x).$$

Indeed, the lemma's integral identity gives $f_k(z) = z^k/k! + o(z^k)$ at zero and shows that $f_k(z)/z^k$ decreases. Conditional on the label-determined J_n, the independent gaps give $\tau_n \sim \mathrm{Gamma}(J_n, 1)$. Hence, for integers $r \geq 1$,

$$\mathbb{E}(\tau_n - J_n)^2 = \mu_n \leq kb_n, \qquad \mathbb{E}(J_n/b_n)^r \leq r\int_0^\infty x^{r-1}\phi_k(x)\,dx < \infty.$$

The bounds are uniform in n. Thus $J_n/b_n \Rightarrow W_1$, and the higher-moment bounds imply convergence of every positive integer moment, in particular $\mu_n/b_n \to m_k$.

Let $D_n = |\mathcal{S}|$ be the number of distinct labels in a block. Unlike in the square case, D_n can differ from $J_n - 1$. The deficit $J_n - D_n$ is at most the number of matching pairs of observations in that block. Since $\{J_n \geq t\}$ is determined before draw t, and draw t matches any given earlier draw with probability $1/n$,

$$\mathbb{E}(J_n - D_n) \leq \frac{1}{n}\sum_{t\geq 1}(t-1)\mathbb{P}(J_n \geq t) = \frac{\mathbb{E}[J_n(J_n-1)]}{2n} = O_k(b_n^2/n) = o(b_n).$$

Consequently, D_n/b_n and $|A|/b_n = (D_n - 1)/b_n$ also converge in distribution to W_1. Relabeling symmetry makes $\mathcal{S}$, and separately A, a uniform subset conditional on its size. Sets from independent blocks remain independent conditional on their respective sizes, but the winner is not independent of its own nonwinning set. Set $\alpha_n = \mathbb{E}|A|/n \sim m_k n^{-1/k}$.

The image at the critical number of maps. For forward composition define

$$I_m = \mathrm{im}(F_m \circ \cdots \circ F_1), \qquad K_m = |I_m|, \qquad \mathcal{B}_m = \mathcal{S}_1 \cap A_2 \cap \cdots \cap A_m.$$

For each deterministic m, IID reversal and absorbing coalescence give $\mathbb{P}(T_n > m) = \mathbb{P}(K_m > 1)$. The forward images satisfy the exact recursion

$$I_m = (I_{m-1} \cap A_m) \cup \begin{cases} \{y_m\}, & I_{m-1} \not\subseteq A_m, \\ \varnothing, & I_{m-1} \subseteq A_m. \end{cases}$$

Inductively $\mathcal{B}_m \subseteq I_m \subseteq \mathcal{B}_m \cup \{y_2, \ldots, y_m\}$. At step k, every element of $I_k \setminus (\mathcal{B}_k \cup \{y_k\})$ is an old winner $y_j \in A_k$, $2 \leq j < k$. For each fixed $j \in \{2, \ldots, k-1\}$, $\mathbb{P}(y_j \in A_k) = \alpha_n$ because the last map is fresh, so the union of these events has probability at most $(k-2)\alpha_n$. The new winner y_k is absent from I_k only if $I_{k-1} \subseteq A_k$. Choosing an anchor in the nonempty I_{k-1}, measurably from the earlier maps, bounds this probability by α_n as well. Outside these events, $I_k = \mathcal{B}_k \dot\cup \{y_k\}$, because $y_k \notin A_k$. Therefore

$$\mathbb{P}\big(K_k - 1 \neq |\mathcal{B}_k|\big) \leq (k-1)\alpha_n = O_k(n^{-1/k}).$$

This accounts for both surviving earlier winners and the event that the new winner is absent from the image.

Condition on $d_1 = |\mathcal{S}_1|$ and $d_j = |A_j|$, $2 \leq j \leq k$. For $Z_n = |\mathcal{B}_k|$, independent uniform subsets give the exact falling-factorial moments

$$\mathbb{E}[(Z_n)_r \mid d_1, \ldots, d_k] = \frac{\prod_{j=1}^{k}(d_j)_r}{(n)_r^{k-1}}, \qquad (u)_r = u(u-1)\cdots(u-r+1).$$

If deterministically $d_j/b_n \to x_j \in (0,\infty)$, these moments tend to $(\prod_j x_j)^r$, since $b_n^k = n^{k-1}$. Bonferroni bounds for $\mathbb{P}(Z_n = 0 \mid d)$ are the even and odd partial sums of the alternating factorial-moment series divided by $r!$. Taking $n \to \infty$ at each fixed truncation, then increasing the truncation, gives $e^{-\prod_j x_j}$. The bounds are uniform for d_j/b_n in a fixed compact subset of $(0,\infty)^k$. Applying the same argument after independent binomial thinning gives the conditional Poisson law, not just its zero probability. The normalized random sizes converge jointly to the independent W_j, and compact truncation therefore yields

$$Z_n \Rightarrow Z, \qquad Z \mid (W_1,\dots,W_k) \sim \mathrm{Poisson}(\Lambda_k), \qquad \mathbb{P}(K_k = 1) \longrightarrow r_k.$$

No moment-determinacy assumption on the mixed law is needed.

Earlier coalescence and the entire later tail. For fixed $1 \le m < k$, condition on the analogous m sizes. The intersection $|\mathcal{B}_m|$ has conditional mean

$$M_{n,m} = \frac{\prod_{j=1}^m d_j}{n^{m-1}} = n^{1-m/k} \prod_{j=1}^m (d_j/b_n) \;\longrightarrow\; \infty \quad \text{in probability.}$$

Two distinct membership indicators in each uniform subset have nonpositive covariance, as do their products across independent subsets. Thus $\mathrm{Var}(|\mathcal{B}_m| \mid d) \le M_{n,m}$. Conditional Chebyshev, followed by truncation where $M_{n,m}$ is small, gives $\mathbb{P}(|\mathcal{B}_m| \le 1) \to 0$. The subset inclusion proves $\mathbb{P}(T_n < k) \to 0$.

For distinct fixed labels, their images under a fresh map differ exactly when at least one belongs to A. For $n \ge 2$ the pair-survival probability is therefore

$$\xi_n = 2\alpha_n - \frac{\mathbb{E}[|A|(|A|-1)]}{n(n-1)} \le 2\alpha_n = O_k(n^{-1/k}).$$

The anchored-image argument of Lemma 9 now gives $\mathbb{E}[K_{m+1} - 1 \mid I_m] \le \xi_n(K_m - 1)$, while that lemma also supplies $\mathbb{E}[K_k - 1] \le k - 2 + k^k$. Thus, for large n,

$$\sum_{m=k+1}^{\infty} \mathbb{P}(T_n > m) \le (k - 2 + k^k)\frac{\xi_n}{1-\xi_n} = O_k(n^{-1/k}).$$

In particular the whole expectation tail vanishes. Together with the previous limits this gives

$$\mathbb{P}(T_n = k) \to r_k, \qquad \mathbb{P}(T_n = k+1) \to 1 - r_k,$$
$$\mathbb{E}T_n = \sum_{m=0}^{k-1} \mathbb{P}(T_n > m) + \mathbb{P}(T_n > k) + \sum_{m=k+1}^{\infty} \mathbb{P}(T_n > m) \longrightarrow k + 1 - r_k.$$

These estimates also give uniform integrability of T_n: eventually $\xi_n \le 1/2$, and the finitely many smaller n have geometric tails because a block is constant with probability $n^{1-k} > 0$.

Requesting block r depends only on preceding completed blocks, independently of its fresh length $J_{n,r}$. Tonelli gives the exact stopped-sum identity

$$\mathbb{E}N_n = \sum_{r \ge 1} \mathbb{E}[J_{n,r}\mathbf{1}\{T_n \ge r\}] = \mu_n \mathbb{E}T_n, \qquad \frac{\mathbb{E}N_n}{b_n} \longrightarrow (k + 1 - r_k)m_k.$$

Constants. Write $W_j = (k!E_j)^{1/k}$ for independent $E_j \sim \mathrm{Exp}(1)$. The change of variables $R = \sum_j E_j$, $U_j = E_j/R$ makes $R \sim \mathrm{Gamma}(k,1)$ independent of the uniform-simplex vector $U \sim \mathrm{Dirichlet}(1,\dots,1)$. Hence

$$r_k = \mathbb{E}_U\left[\left(1 + k!\Big(\prod_j U_j\Big)^{1/k}\right)^{-k}\right]$$
$$= (k-1)! \int_{\substack{u_1,\dots,u_{k-1} \ge 0 \\ \sum_{j<k} u_j \le 1}} \frac{du_1 \cdots du_{k-1}}{\left[1 + k!\big(\prod_{j=1}^k u_j\big)^{1/k}\right]^k}, \qquad u_k = 1 - \sum_{j<k} u_j.$$

Using $\min_j E_j \le (\prod_j E_j)^{1/k} \le k^{-1}\sum_j E_j$ and the Laplace transforms of $\mathrm{Exp}(k)$ and $\mathrm{Gamma}(k,1)$ gives

$$[1+(k-1)!]^{-k} \le r_k \le [1+(k-1)!]^{-1}.$$

For the square case the constant evaluates explicitly:

$$r_2 = \int_0^\infty \int_0^\infty xye^{-(x+y)^2/2}\,dx\,dy = \frac{1}{6}\int_0^\infty u^3 e^{-u^2/2}\,du = \frac{1}{3}.$$

This recovers $\mathbb{E}T_n \to 8/3$ and $\mathbb{E}N_n/\sqrt{n} \to (8/3)\sqrt{\pi/2}$. □

The bounds imply $r_k \to 0$, so Stirling's formula gives $(k+1-r_k)m_k \sim k^2/e$ as $k \to \infty$ *after taking the* $n \to \infty$ *limit at each fixed* k. This limit gives no joint asymptotic equivalent for $k = k(n)$ or uniform approximation rate. These constants describe this sampler on exactly uniform inputs and need not be optimal among all universal exact samplers.

B.3 Finite uniform sources and joint growth

An exact finite-size recursion. On an exactly uniform positive support of size n, let D be the number of distinct labels in one collision block, and write $\pi_d = \mathbb{P}(D = d)$. Conditional on $D = d$, the nonwinning set A is uniform of size $d-1$. Its complement B within the positive support has size $b = n - d + 1$. All of B maps to the winner in B, and all other labels are fixed. For an independent current image I of size m, let $R = |I \cap B|$ count its values that map to the winner. If $R \ge 1$, those R values merge into one; if $R = 0$, every value stays fixed. A fresh map therefore gives

$$K' = m - \max(R-1, 0), \qquad R \mid D = d \sim \mathrm{Hypergeom}(n, m, n-d+1).$$

The resulting nonincreasing cardinal process is Markov: the fresh block is independent, and this transition depends only on m. Its kernel is

$$P_{m\ell} = \sum_{d=1}^{n} \pi_d \sum_{r=0}^{\min(m,b)} \frac{\binom{m}{r}\binom{n-m}{b-r}}{\binom{n}{b}} \mathbf{1}\{\ell = m - \max(r-1,0)\}, \qquad b = n-d+1,$$

where out-of-range binomial coefficients are zero. It uses the original n-label oracle at every state m, rather than replacing it by an m-label source. First-step analysis gives the exact remaining-map means

$$E_1 = 0, \qquad E_m = \frac{1 + \sum_{\ell<m} P_{m\ell}E_\ell}{1 - P_{mm}} \quad (2 \le m \le n).$$

The denominator is positive because a block with $D = 1$ is constant and has probability n^{1-k}. For $n \ge 2$, fixed-time reversal and monotone constancy, as in Appendix A.6, give $\mathbb{E}T = E_n$. For $n = 1$, the actual algorithm still generates its first map, so $T = 1$ and $N = k$. The forced-first-map identity $\mathbb{E}T = 1 + \sum_d \pi_d E_d$ handles both cases. Zero-mass ambient labels add no image values after the first map and do not change these conclusions.

Both input quantities of this recursion have finite coefficient formulas, where $[z^s]$ denotes the coefficient of z^s:

$$\pi_d = n\binom{n-1}{d-1} \sum_{s=d-1}^{(d-1)(k-1)} \frac{(k+s-1)!}{(k-1)!\,n^{k+s}} [z^s] \left(\sum_{r=1}^{k-1} \frac{z^r}{r!}\right)^{d-1},$$

$$\mu_k = \sum_{j=0}^{n(k-1)} j! n^{-j} [z^j] \left(\sum_{r=0}^{k-1} \frac{z^r}{r!}\right)^n.$$

For the first, choose the winner and the $d-1$ other observed labels. If their positive counts sum to s, the legal prefix before the final winner has $(k+s-1)!/((k-1)!\prod r_i!)$ orderings. For $d = 1$, the polynomial is one and only $s = 0$ contributes. For the second, the coefficient is the multinomial probability of no k-fold occurrence after j draws, and summing these survival probabilities gives μ_k. Predictable block requests again give $\mathbb{E}N = \mu_k \mathbb{E}T$.

For example, the recursion on the exactly uniform three-label source at $k = 2$ gives $\mu_2 = 26/9$, $\mathbb{E}T = 225/91$, and $\mathbb{E}N = 50/7$. The reduction to image cardinality uses exchangeability of the uniform source. For a nonuniform law, the labels in the image, rather than only their number, affect the next transition.

A joint-growth obstruction. The fixed-k uniform limit must not be read as a guarantee when k grows with the support size. A pair of tracked labels gives a stronger construction-specific boundary than the binary example.

Proposition 14 (A support-uniform constant must grow faster than 2^k)**.** *Let $T_{n,k}$ be the map count of Algorithm 1 on the exactly uniform law on n labels. Put*

$$n_k = \lceil e^{9k/25} \rceil, \qquad b = \frac{8}{3} e^{-1/4} > 2.$$

Then

$$\liminf_{k\to\infty} \frac{\mathbb{E}T_{n_k,k}}{b^k} \geq \frac{1}{2}.$$

The same lower bound holds for any C_k in $\mathbb{E}N \leq C_k \mu_k$ that is valid for this construction uniformly over support sizes, for example on a fixed countably infinite ambient universe. It is not a statement for each fixed finite alphabet, nor a lower bound on the cost ratio to the optimal universal exact sampler.

Proof. Poissonize the n categories at rate one each, which preserves their uniform arrival-label sequence. Write

$$g_k(t) = e^{-t} \frac{t^{k-1}}{(k-1)!}, \qquad \phi_k(t) = e^{-t} \sum_{r=0}^{k-1} \frac{t^r}{r!}.$$

Here g_k is the density of a label's kth arrival time. For distinct labels x, y, let $\delta_{n,k} = \mathbb{P}(F(x) = F(y))$. Either x wins with y unseen, y wins with x unseen, or another label wins with both unseen. Independence gives, for $n \geq 3$,

$$\begin{aligned}\delta_{n,k} = 2 \int_0^\infty g_k(t) e^{-t} \phi_k(t)^{n-2}\, dt \\ + (n-2) \int_0^\infty g_k(t) e^{-2t} \phi_k(t)^{n-3}\, dt.\end{aligned}$$

By exchangeability this probability is the same for every distinct pair. For independent forward maps, a tracked pair consequently has constant conditional merger probability $\delta_{n,k}$ before merging. Its merger time is geometric, and global coalescence cannot occur earlier. Fixed-time reversal transfers the global absorption tails to the backward algorithm, so $\mathbb{E}T_{n,k} \geq 1/\delta_{n,k}$.

Set $a = 3/8$, $c = 9/25$, and $I(u) = u - 1 - \log u$. Split the first integral at ak and drop the second integral's ϕ factor. Since ϕ decreases,

$$\delta_{n,k} \leq 2^{1-k} \mathbb{P}\{\mathrm{Gamma}(k,1) \leq 2ak\} + 2^{1-k} \phi_k(ak)^{n-2} + (n-2)3^{-k},$$

where the gamma law has rate one. Its lower-tail Chernoff bound gives $\mathbb{P}\{\mathrm{Gamma}(k,1) \leq (3/4)k\} \leq e^{-kI(3/4)}$. Thus the first term is at most $2e^{-\gamma k}$, with $\gamma = \log(8/3) - 1/4$.

For the middle term, the Poisson mass at k and the elementary integral bound $k! \leq ek(k/e)^k$ imply

$$1 - \phi_k(ak) \geq e^{-ak} \frac{(ak)^k}{k!} \geq \frac{e^{-kI(a)}}{ek}.$$

For $n = n_k$ and $k \geq 4$, $n - 2 \geq e^{ck}/2$. Hence

$$\phi_k(ak)^{n-2} \leq \exp\left(-\frac{e^{\beta k}}{2ek}\right), \qquad \beta = \frac{197}{200} - \log\frac{8}{3} > 0.$$

This is smaller than every exponential in k. Finally, $(n-2)3^{-k} \leq e^{-\eta k}$, where $\eta = \log 3 - 9/25 > \gamma$. Therefore $\delta_{n_k,k} \leq (2 + o(1))e^{-\gamma k}$, proving the claim because $b = e^\gamma$. The signs are strict without numerical approximation: $\sum_{r=0}^4 (197/200)^r/r! > 8/3$, $\log(9/8) > 15/128 > 11/100$, and $\log(4/3) > 1/4$. □

This obstruction concerns the ratio to the first-block mean μ_k for this algorithm. The output-multiplicity lower bound does not identify μ_k with the optimal exact-sampler cost. Nor does the known-binary walk in Appendix B.1 provide one sampler for arbitrary unknown countable support. Determining the sharp growth rate of the worst-case coalescence factor remains open.

C Connections to exact distribution transformations

C.1 Connections to other exact-sampling frameworks

To place our result among exact distribution transformations, we compare what information each method uses, which distribution it targets, and what it guarantees. Methods for detecting coalescence in CFTP are established (Propp & Wilson, 1996; 1998; Murdoch & Green, 1998; Huber, 2004); here the task is to construct maps from samples and bound their coalescence cost independently of the input. Perfect sampling includes methods with different coupling and execution requirements. Fill (1998) and Fill et al. (2000) develop rejection-based alternatives, while read-once CFTP avoids repeated access to the same random inputs (Wilson, 2000). Dominated CFTP handles spatial point processes using a dominating process (Kendall & Møller, 2000). The relationship between backward coupling and uniform ergodicity is studied by Foss & Tweedie (1998). Our algorithm uses independent collision blocks and an explicit finite-image representation, and its cost bound is proved directly for that representation, rather than inferred from a generic perfect-sampling guarantee.

The choice of a fixed integer power is also substantive. Bernoulli factories can simulate some non-integer power functions using power-series constructions (Mendo, 2019), whereas finite-automata restrictions change which multivariate rational transformations are possible (Paes Leme & Schneider, 2026). Those computational models provide useful comparisons, but a noninteger power has no direct analogue of the stopping rule that waits for a label's kth occurrence.

There is a related distinction between streaming frequency sampling and transforming an unknown source distribution. Frequency-moment sketches summarize empirical count vectors (Alon et al., 1999), and L_p samplers target coordinate probabilities proportional to powers of those counts (Jayaram & Woodruff, 2021; Jayaram et al., 2022). Even exact sampling from a finite empirical frequency vector does not in general give an exact sample from the corresponding power of the population probabilities. Our target is the latter, and the source-sample count is the resource being bounded. At the level of scalar estimation, collision statistics govern the estimation of power sums and Rényi entropy (Obremski & Skorski, 2017; Crouch et al., 2016). Estimating integer-order Rényi entropy on n labels needs order $n^{1-1/k}$ samples (Acharya et al., 2017). Our result identifies the necessary and sufficient scale for exact powering on each input as $1/\|p\|_k$.

Bernoulli races permit resampling with intractable particle weights in sequential Monte Carlo (Schmon et al., 2019). They share the principle of replacing unavailable probabilities with random experiments. Here the experiments are whole collision blocks, and the output distribution is determined by a reversible random-map kernel.

C.2 Rejection based on identical observations

CALM's exact likelihood-free sampler at temperature $1/k$ draws k samples at once and restarts unless all are identical, so its expected cost is k/S_k (Shao et al., 2025, Algorithm 1 and Theorem 2). Our comparisons use the stronger early-abort version, which draws each group one sample at a time and abandons it at the first mismatch with its initial label. A successful group has k copies of label i with probability p_i^k, so the accepted label has law q_k. With $S_j = \sum_i p_i^j$, the mean trial length is $2 + \sum_{j=2}^{k-1} S_j$, and the success probability is S_k. The renewal equation for the expected total cost $\mathbb{E}N_{\mathrm{rej}}$ gives

$$\mathbb{E}N_{\mathrm{rej}} = 2 + \sum_{j=2}^{k-1} S_j + (1 - S_k)\,\mathbb{E}N_{\mathrm{rej}}, \qquad \mathbb{E}N_{\mathrm{rej}} = \frac{2 + \sum_{j=2}^{k-1} S_j}{S_k}.$$

The sum is empty at $k = 2$. Uniform inputs give $\mathbb{E}N_{\mathrm{rej}} = \Theta_k(n^{k-1})$. This comparison is between algorithms with the same target and sample interface, and it illustrates the benefit of using matches across observations that rejection would discard.

C.3 Sampling interfaces and applications

A powered marginal differs from a powered distribution over latent paths. If $Y = g(\omega)$ is a fixed discrete observable of a random execution ω, then $p_i = \mathbb{P}(g(\omega) = i)$ sums over the preimage

$g^{-1}(i)$. Applying the present algorithm to independent $g(\omega)$ samples powers these sums. Powering the path probabilities before applying g generally gives a different law. This distinction is relevant to probabilistic programming and inference over generated text (Dohan et al., 2022; Lew et al., 2023; Zhao et al., 2024). Powering a marginal posterior that cannot be evaluated is also used for marginal maximum a posteriori estimation (Doucet et al., 2002).

For language models, full-sequence power samplers use Markov chain, sequential Monte Carlo, or autoregressive constructions (Karan & Du, 2026; Azizi et al., 2026; Ji et al., 2026; Zhou et al., 2026). Marginal sharpening explicitly targets aggregated answers (Arzhantsev et al., 2026). CALM has the closest access model to ours (Shao et al., 2025). Its generative head is likelihood-free, so CALM develops temperature sampling from a black-box sampler, with an exact group-rejection stage for the integer part of the inverse temperature and a Bernoulli-factory stage for the fractional part (Mendo, 2019). Noting that agreement becomes very unlikely at low temperatures, CALM concludes that "a more sample-efficient approximate algorithm is therefore needed" (Shao et al., 2025, Section 5.2) and proposes a batch approximation. The approximation draws a batch of samples and chooses among labels seen at least k times, with weights equal to their numbers of matching k-subsets, falling back to smaller multiplicities when no label reaches k. CALM proves that it is consistent as the batch size grows and notes that it is biased for every finite batch. The batch method already exploits matches across samples, the idea behind our construction. Our sampler keeps exactness at the collision scale for temperatures $1/k$ with integer k, and it could replace the rejection sampler in the integer stage of CALM's scheme. Self-consistency and semantic uncertainty also operate on aggregated outcomes, but have selection or uncertainty objectives rather than the exact transformed law studied here (Wang et al., 2023; Chen et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). Aggregating sampled answers also underlies inference-time computation (Snell et al., 2025).

The sampling interface is also different from score or density access. Discrete diffusion language models typically provide likelihoods only through variational bounds or Monte Carlo estimates of them (Lou et al., 2024; Sahoo et al., 2024; Nie et al., 2025); a method using only their samples avoids these estimates. High-accuracy diffusion samplers use score information to control approximation error (Chen et al., 2026), whereas our primitive takes independent discrete outputs and produces an exactly specified transformation. Input-specific promises, likelihood evaluations, and amortization across multiple outputs each change the comparison class in Theorem 2.

D Formal verification in Lean 4

Every theorem, lemma, and proposition in this paper is formally verified in Lean 4 with Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020). The verified results fall into four groups.

- **Exactness and cost.** Theorem 1, including the square constant $C_2 = 691/64$ (Lemma 11) and the bounds on dictionary operations and storage. Its proof uses the collision-time bounds (Lemma 5), detailed balance (Lemma 6), the finite representation (Lemma 7), pair contraction (Lemma 8), and the image and coalescence bounds (Lemmas 3, 9 and 10).
- **Lower bound.** Theorem 2 for universal exact samplers, and the fact that Algorithm 1 is one.
- **Uniform inputs.** Propositions 4 and 13, namely the limiting law of the number of maps, its mean, and the sample-cost coefficient.
- **Dependence on the power.** Propositions 12 and 14.

Several unnumbered calculations are also verified, including the bias of the first collision (Section 4.1), the finite uniform recursion (Appendix B.3), the cost of early-abort rejection (Appendix C.2), and the fair-coin costs (Appendix B.1). The Lean development contains no unfinished proofs (`sorry`) and uses only Lean's standard axioms `propext`, `Classical.choice`, and `Quot.sound`.