Temperature Sampling is Entropy-Optimal: An Information-Theoretic Framework for LLM Decoding

Jin Sima Nikolaos Papagiannis Ananth Grama Wojciech Szpankowski
Purdue University, West Lafayette, IN 47907

Abstract

Degeneracy in large language models, which is manifested as repetitive, low-quality output, remains a persistent failure mode in autoregressive generation. Although heuristics such as temperature scaling, nucleus sampling, and top-k sampling are widely used in practice, their effectiveness lacks a unifying theoretical foundation. Furthermore, these techniques are inherently static and do not offer adaptation to evolving input distributions. In this paper, we develop a rigorous information-theoretic framework for non-repetitive sequence generation and use it to provide theoretical underpinnings for temperature-based sampling. We formalize the avoidance of repetition through the lens of recurrence time: leveraging Kac's Lemma, we adopt maximization of the per-token average entropy under a global cross-entropy constraint as the principled objective for avoiding degeneracy. The solution to the resulting constrained entropy-maximization problem is exactly temperature sampling with a fixed offline optimal temperature. This provides the first clean theoretical justification for this widely used heuristic. Our goal is not to reinterpret temperature sampling, rather, we show that it is the unique optimizer of a principled objective. We then extend this framework to the practically relevant and novel online setting, where the model's token distributions evolve and are revealed sequentially. We formulate the corresponding online entropy-maximization problem, characterize its optimal solution, and design the Entropic Online Sampling (EOS) algorithm, a computationally efficient adaptive temperature-control scheme. We establish that EOS is asymptotically optimal: its cumulative entropy loss relative to the offline optimum is $o(n)$ with high probability under mild conditions, and we prove a matching lower bound.

1 Introduction

Despite the empirical success of large language models (LLMs) across a wide range of tasks [1, 14], fundamental theoretical questions about their decoding procedures remain unresolved. A central and persistent failure mode is degenerate generation: the tendency of models to produce repetitive, low-entropy, or otherwise vacuous output sequences. Standard procedures such as greedy search and beam search are known to exacerbate this phenomenon [8, 11], and the remedies widely adopted in practice, such as temperature scaling, nucleus sampling, top-k truncation, and repetition penalties are heuristics, adopted for empirical utility rather than derived from a principled optimization criterion. A fundamental mathematical treatment of the problem remains to be established.

The central thesis of this paper is that non-repetitive, expressive sequence generation admits a clean formulation as a constrained entropy maximization problem. In the offline setting, where token distributions are known in advance, this formulation yields temperature sampling as its exact, analytically derived solution, providing a rigorous theoretical justification for this widely used heuristic. In the more challenging online setting, in which the distributions are revealed sequentially, the problem naturally connects to the online convex optimization framework [2, 12]. However, the presence of a global constraint coupling cross-entropy across all time steps makes the problem structurally distinct from standard online learning, and requires new analytical solutions, which we develop in our work. The resulting Entropic Online Sampling (EOS)

Algorithm 1 proposed in this paper achieves an entropy loss of $o(n)$ relative to the offline optimum, with a matching lower bound, proving that this rate cannot be improved under mild conditions on token distributions. To our knowledge, this is the first rigorous formulation and justification of the online constrained entropy maximization problem.

We view a generative model as an autoregressive sequence generator. At each discrete time instant t, a conditional probability distribution for token w, $p_t := P(w \mid w^{t-1})$, is computed by the generative model based on previously sampled tokens $w^{t-1} := (w_1, \ldots, w_{t-1})$. This conditional distribution is used to construct a sampling distribution $q_t(w)$, from which the next token w_t is selected. Our goal is to construct the sampling distribution that yields optimal generation diversity.

In addition to high likelihood, generative models must also optimize for expressiveness, or equivalently, the avoidance of repetition. Informally, a desirable generative model is one in which a typical token sequence w^n does not repeat too quickly [9, 12]. A natural way to formalize this is through the recurrence time (or return time) $S(w^n)$, the average number of tokens until the sequence returns to its original state, which we would like to be as large as possible. Classical information theory provides a clear intuition for what makes recurrence times long: for stationary ergodic processes W, Kac's Lemma [9] states that $S(w^n) \sim 1/P(w^n)$, and for typical sequences the asymptotic equipartition property (AEP) [3] gives $P(w^n) \sim 2^{-nH(W)}$, where $H(W)$ is the entropy rate of the generating process. Combining the two yields $S(w^n) \sim 2^{nH(W)}$, suggesting that high-entropy sampling distributions induce long recurrence times. Our setting is non-stationary—token distributions q_t vary across time steps, and dependence on past samples is treated as exogenous (see below)—so this argument does not apply directly as a derivation. Instead, we use it as motivation: it identifies entropy maximization as a principled target for avoiding degeneracy, and we then adopt the per-token average entropy under a global cross-entropy constraint as the formal objective and analyze it on its own terms.

In our setting, distribution W corresponds to the sampling distribution q_t. Our objective is therefore to maximize the entropy $H(q_t)$, more precisely, $\sum_{1\le t\le n} H(q_t)$ over q_t. At the same time, the sampling distribution should produce "reasonable tokens" w^n whose probability $P(w^n)$ is above a prescribed threshold (improving downstream performance). In practice, the probability $P(w^n)$ of a sampled token sequence decays exponentially with sequence length n. It is desirable that the exponent be upper bounded by some γ to avoid tokens with very small probability, as implemented in current AI systems. Hence, the probability of a "reasonable" token sequence is lower bounded by the threshold $2^{-n\gamma}$. In Theorem 3.1, we prove that this condition is equivalent to

$$-\sum_{t\in[n]}\sum_{w} q_t(w)\log P\big(w \mid w^{t-1}\big) \le \gamma n.$$

In Theorem 3.2, we show that the problem of maximizing the entropy of the generative process under the probability threshold constraint yields an optimal sampling distribution for offline optimization of the form of temperature sampling:

$$q_t(w) = \frac{P^{\lambda^*}\big(w \mid w^{t-1}\big)}{\sum_{w'} P^{\lambda^*}(w' \mid w^{t-1})},$$

where $\lambda^* = 1/T^*$ is the inverse of an optimal temperature T^*.

Temperature is typically chosen empirically (e.g., $T = 0.7$), without any formal underpinnings or notions of optimality. Furthermore, existing offline-trained LLMs are unable to adapt their sampling distributions in real time in response to distributional shifts or evolving contexts. This lack of adaptability limits their effectiveness in applications that require timely, context-aware, and diverse generation. To address these challenges, we introduce an online formulation of the constrained entropy maximization problem. Informally, in the online entropy optimization problem, we maximize the total entropy $\sum_{1\le t\le n} H(q_t)$ over q_t that depends only on the probability distributions $\{p_i\}_{i=1}^{t}$ observed until time t, subject to the probability threshold constraint $w^n \in \mathcal{W}_\gamma^n := \{w^n : P(w^n) \ge 2^{-n\gamma}\}$ for $\gamma > 0$, and $\mathcal{W}^n$ being the set of all sequences

of tokens of length n. Please see (1)–(2) in Section 2 for a formal statement of this online optimization problem.

Our Contributions. Our contributions are as follows: *(a) Theoretical foundations of temperature sampling.* We show that temperature sampling is the exact solution to the offline constrained optimization problem (Theorem 3.2), grounded in the information-theoretic concepts of recurrence time and entropy rate via Kac's Lemma [9] and the AEP [3]. *(b) Online problem and the EOS algorithm.* We fully characterize the optimal online solution (Lemma 3.3) and design the Entropic Online Sampling (EOS) algorithm (Algorithm 1), a computationally efficient adaptive scheme. The online optimum corresponds to a time-varying inverse temperature λ_t^{on}; EOS tracks it with entropy loss $O(R \log n)$ relative to the offline optimum (Theorem 3.5), where $R = \max_{t \in [n]} R_t$ measures cumulative constraint deviation. *(c) Matching lower bound and asymptotic optimality.* We prove that no online algorithm can achieve entropy loss $o(R)$ (Theorem 3.4), establishing that EOS is tight up to logarithmic factors. Under mild stationarity conditions on $\{p_t\}$, we show $R = O(\sqrt{n \log(1/\delta)})$ with probability $1 - \delta$ (Theorem 3.8), so the entropy loss of EOS is $o(n)$, yielding asymptotic optimality.

Related Literature. The observation that maximum-likelihood decoding leads to degenerate outputs is well-documented. Holtzman et al. [8] identified nucleus sampling as an empirical fix for the repetition and incoherence produced by greedy and beam-search decoding; Meister and Cotterell [11] offered a post-hoc information-theoretic reinterpretation of beam search, arguing that it implicitly targets a uniform information-density objective. Temperature scaling, top-k sampling [4], and repetition penalties [10] are similarly motivated by intuition rather than derived from an optimization problem. A related line of work on unlikelihood training [15] targets repetition at training time, but still relies on static, offline objectives. None of these approaches derives an optimal decoding rule from first principles, nor do they provide performance guarantees relative to well-defined benchmarks. Our work fills this gap by identifying entropy maximization as the canonical objective and recovering temperature sampling as its exact solution.

Information theory has been used to analyze language models in several ways: calibration and uncertainty estimation [7, 13], the mismatch between likelihood-based training and sequence-level generation quality [11], and the connection between entropy collapse and model degradation [8]. However, these analyses are largely descriptive in that they characterize failure modes rather than prescribe solutions. Our work is constructive in that we formulate a mathematically precise optimization problem whose solution is a concrete, implementable decoding algorithm, and we derive all performance guarantees from the structure of the problem.

The theory of online convex optimization [2, 12] and mirror descent provides a rich toolkit for sequential decision-making under non-stationarity, with tight regret bounds in adversarial settings. More recent work on online universal learning [16] extends these ideas to information-theoretic objectives. Integration of online learning into autoregressive generation, however, has received limited theoretical treatment [6] and prior work lacks guarantees on entropy or diversity. Our formulation inherits the sequential structure of online convex optimization but differs in the presence of a global, coupled constraint (the probability threshold across the entire sequence), which requires new analytical tools beyond standard regret analysis.

Studies on dataset shift and predictive uncertainty [7, 13] have established that large models degrade under distributional change, but focus on predictive accuracy rather than generative diversity. Our work addresses a complementary concern: maintaining entropy and expressiveness as token distributions evolve, with provable guarantees under mild assumptions.

The remainder of the paper is organized as follows. Section 2 formalizes the online and offline optimization problems. Section 3 presents the main theoretical results. Proofs of all results are presented in the appendices.

2 Problem Setup

We view the model's estimate of the probability distribution over token sequences $w^n \in \mathcal{W}^n$ as $P(w^n) = \prod_{t=1}^n P(w_t \mid w^{t-1}) = \prod_{t=1}^n p_t(w_t)$, obtained from the conditional distributions $P(w_t \mid w^{t-1})$, $t \in [n] = \{1, \ldots, n\}$ given by the model. Let $\mathcal{W}_\gamma^n = \{w^n : P(w^n) \geq 2^{-n\gamma}\}$ be the set of length-n sequences with probability at least $2^{-n\gamma}$. The set $\mathcal{W}_\gamma^n$ represents sequences that are considered reasonable by the model, that is, rare tokens are eliminated since they may lead to poor outputs downstream.

The goal is to sample sequences w^n in a sequential manner such that w^n is, with high probability, in $\mathcal{W}_\gamma^n$. Specifically, w_t is sampled sequentially for $t \in [n]$ according to a sampling distribution $q_t = \phi_t(\{p_i\}_{i=1}^t) \in \Delta(\mathcal{W})$, which is a function of prior distributions p_i known up to time t. Therefore, the distribution of a sampled sequence w^n is $\prod_{t=1}^n q_t(w_t)$. When $q_t = p_t$, w^n has distribution $P(w^n)$. Note that widely used sampling algorithms, such as top-k, top-p, or temperature sampling, are specific examples of q_t. In general, the conditional distribution $p_t(w) = P(w \mid w^{t-1})$ depends on the sampled tokens w^{t-1} in the past. However, since the behavior of such dependence is hard to characterize and the effect of the choice of token w_t on future conditional distributions $P(w_{t'} \mid w^{t'-1})$, $t' \geq t$, is unpredictable, we treat $p_t(w) = P(w \mid w^{t-1})$ as arbitrarily given.

One strategy to sample sequences with high probability is to sample every token with the maximum conditional probability, i.e., $q_t(w) = 1$ if $w = \arg\max_{w \in \mathcal{W}} p_t(w)$ and 0 otherwise. However, this strategy generates deterministic sequences, which reflects a lack of expressiveness or creativity, and results in degenerate (repeated) sequences. We are interested in the tradeoff between the probability guarantee γ (which is the measure of goodness of generated sequences considered in this paper) and the creativity/expressiveness of the sampled sequence, which is measured by the entropy of the sampling probability distributions q_t, $t \in [n]$, and is related to the repetitiveness of the sampled sequence. Specifically, we define the following online optimization problem

$$\max_{q_t = \phi_t(\{p_i\}_{i=1}^t) \in \Delta(\mathcal{W}): t \in [n]} \quad \sum_{t=1}^n H(q_t) \tag{1}$$

$$\text{s.t.} \quad -\sum_{t=1}^n \sum_{w \in \mathcal{W}} q_t(w) \log p_t(w) \leq \gamma n, \tag{2}$$

where γ is a given parameter. We justify below that a solution to the above optimization problem achieves the creativity-probability tradeoff considered in the paper. We will also provide a theoretical justification for the commonly used temperature sampling as the solution to the above online optimization problem.

Note that Problem (1) corresponds to online optimization of q_t, $t \in [n]$ in the sense that no knowledge of p_i, $i > t$ is assumed. Consider, in contrast, the following offline (static) problem, where $\{q_t\}_{t=1}^n$ are optimized with $\{p_t\}_{t=1}^n$ known:

$$\max_{q_t \in \Delta(\mathcal{W}): t \in [n]} \quad \sum_{t=1}^n H(q_t)$$

$$\text{s.t.} \quad -\sum_{t=1}^n \sum_{w \in \mathcal{W}} q_t(w) \log p_t(w) \leq \gamma n. \tag{3}$$

In Theorem 3.2, we show that the optimal offline solution to (3) is temperature sampling with fixed temperature. However, this may not hold for the online setting, as shown in Lemma 3.3. In Algorithm 1 we present an online learning algorithm for (1) and prove in Theorems 3.4 and 3.5 that the achieved entropy is less than the optimal entropy for the offline problem (3) by at most $o(n)$ with high probability (whp) under some mild conditions on p_t.

3 Main Results

In this section, we present our main theoretical results. We first show in Theorem 3.1 that the probability threshold constraint for "reasonable words" $\mathcal{W}_\gamma^n$ is asymptotically equivalent to our condition (2). In Theorem 3.2, we address the offline optimal solution, which is hard to implement in real-world settings. We then present an efficient online algorithm (Algorithm 1) that is asymptotically optimal under mild assumptions (Theorem 3.5) and efficient to implement.

3.1 Preliminary Results

We start by establishing three preliminary results that highlight key aspects of our online optimization. The first result proves the equivalence of the constraint (2) and the set of reasonable words $\mathcal{W}_\gamma^n$.

Theorem 3.1. *Let w^n be the sequence sampled by q_t, $t \in [n]$ sequentially, i.e., the probability of sampling w is $\prod_{t\in[n]} q_t(w_t)$. If*

$$\sum_{t\in[n]}\sum_{w\in\mathcal{W}} q_t(w)\log p_t(w) + \gamma n \;\geq\; -O\left(\sqrt{n\log\tfrac{1}{\delta}}\right), \tag{4}$$

then

$$\Pr\left(P(w^n) \geq 2^{-\gamma n - O(\sqrt{n\log(1/\delta)})}\right) \;\geq\; 1-\delta. \tag{5}$$

Conversely, if (5) holds and $p_t(w) \geq \rho$ for $w \in \mathcal{W}$ and some constant $\rho > 0$, then

$$\sum_{t\in[n]}\sum_{w\in\mathcal{W}} q_t(w)\log p_t(w) + (1-\delta)\gamma n \;\geq\; -O\left(\sqrt{n\log\tfrac{1}{\delta}}\right) - \delta n(\log\rho^{-1}+\gamma). \tag{6}$$

By choosing $\delta = O(1/\sqrt{n})$, the constraint $w^n \in \mathcal{W}_\gamma^n$ is asymptotically equivalent to (2) with high probability.

The next two results present solutions to the offline and online optimization problems, respectively.

Theorem 3.2. *The optimal solution for the offline optimization problem (3) is given by*

$$q_t(w) = Q_t^{\text{off}}(w) = \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}, \tag{7}$$

where λ^* is the unique solution to:

$$-\sum_{t\in[n]}\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}\log p_t(w) = \gamma n. \tag{8}$$

Lemma 3.3. *The optimal solution to the online problem (1) and (2) is given by:*

$$q_t(w) = Q_t^{\text{on}} = \frac{p_t^{\lambda_t^{\text{on}}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{\text{on}}}(w)}, \qquad t \in [n] \tag{9}$$

for some λ_t^{on}, $t \in [n]$, such that:

$$-\sum_{t\in[n]}\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda_t^{\text{on}}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{\text{on}}}(w)}\log p_t(w) = \gamma n. \tag{10}$$

Proof. Following similar arguments in the proof of Theorem 3.2, we find that $q_t(w)$ of the form

$$q_t(w) = \frac{p_t^{\lambda_t^{\text{on}}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{\text{on}}}(w)}, \qquad t \in [n],$$

maximizes $H(q_t)$ for fixed value of $\sum_{w\in\mathcal{W}} q_t(w)\log p_t(w) + \gamma$. Hence (9) follows. To show that (10) holds, suppose on the contrary the equality does not hold, we have that

$$-\sum_{t\in[n]}\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda_t^{\text{on}}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{\text{on}}}(w)} \log p_t(w) < \gamma n.$$

Note that the left hand side decreases with λ_t, $t \in [n]$. One can decrease λ_n^{on} so that

$$-\sum_{t\in[n]}\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda_t^{\text{on}}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{\text{on}}}(w)} \log p_t(w) \le \gamma n$$

still holds. Note that $H(Q_n^{\text{on}})$ decreases with λ_n^{on}. Hence, decreasing λ_n^{on} increases $H(Q_n^{\text{on}})$, contradicting the optimality of Q_t^{on}, $t \in [n]$. Hence, we have (10). □

3.2 An Efficient Online Algorithm

Our solution to the online optimization problem is not algorithmically efficient and hard to realize in practice (i.e., in an online setting). To handle this, in the next two main results, we first establish a lower bound on the entropy loss for any online algorithm, when compared to the offline algorithm. We then design an approximate, efficient online Algorithm 1 and show in Corollary 3.6 that it achieves asymptotically online optimality, under mild conditions.

We start with a general lower bound proved in Appendix A. Throughout the rest of the paper we will assume that $p_t(w) \ge \rho$ for some $\rho > 0$.

Theorem 3.4 (Lower Bound)**.** *For any online algorithm $\mathcal{A}$ that selects q_t as a function of $\{p_i\}_{i=1}^t$ and satisfies (2), let $q_t^{\mathcal{A}}$, $t \in [n]$, be the output of any algorithm $\mathcal{A}$. Then for any γ and $R \le \gamma n/6$ satisfying:*

$$-\frac{(|\mathcal{W}|-1)\log\left(\frac{1-\exp(-\gamma/3)}{|\mathcal{W}|-1}\right)}{|\mathcal{W}|} + \frac{\gamma}{3|\mathcal{W}|} > \frac{4\gamma}{3}$$

and $\log|\mathcal{W}| \ge \frac{4\gamma}{3}$, there exists a set of distributions p_t, $t \in [n]$, such that:

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \;\ge\; \Omega(R(n)), \tag{11}$$

where $R := R(n) := \max_{t\in[n]} R_t$ with R_t being

$$R_t = \sum_{i=1}^{t}\sum_{w\in\mathcal{W}} Q_i^{\text{off}}(w)\log p_i(w) + \gamma t \tag{12}$$

for $t \in [n]$.

Sketch of the proof. We briefly describe the main idea of the proof leaving details to Appendix A. In addition to $R_t =: R_t^{\text{off}}$ defined above, we introduce $R_t^{\mathcal{A}} = \sum_{i=1}^{t} \sum_{w \in \mathcal{W}} q_i^{\mathcal{A}}(w) \log p_i(w) + \gamma t$. Next we construct two distributions $p_{t,1}$ and $p_{t,2}$ such that they are equal up to time $0 \le t \le \alpha = 3R/\gamma$ but different otherwise (see (24)). We will prove that $R_\alpha^{\text{off}} = -R$ for $p_t = p_{t,1}$ and $R_\alpha^{\text{off}} = R$ for $p_t = p_{t,2}$, which implies $|R_\alpha^{\text{off}} - R_\alpha^{\mathcal{A}}| = \Omega(R)$ either when $p_t = p_{t,1}$ or $p_t = p_{t,2}$. Finally, after some algebraic manipulations, we show that

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) = \Omega\Big(|R_\alpha^{\text{off}} - R_\alpha^{\mathcal{A}}|\Big) = \Omega(R)$$

as needed. □

Algorithm 1: Solving λ_t

Input: γ, n, p_t at time t.

Output: q_t at time t.

for $t = 1$ **to** n **do**

Let $q_t^{\text{on}}(w) = \dfrac{p_t^{\lambda_t}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda_t}(w)}$, where λ_t satisfies

$$-\sum_{i \in [t]} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda_t}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda_t}(w)} \log p_i(w) = \gamma t. \tag{13}$$

The lower bound tells us that the best achievable performance for any algorithm is bounded by $R(n)$ away from the offline optimal, which can go up to $O(n)$. However, as we prove in Theorem 3.8 below, there exist a large class of p_t for which $R(n) = o(n)$.

To complete the picture, we need an upper bound on the entropy loss of $o(n)$. That is, we need an online algorithm that approximates the optimal entropy for the offline setting well. This algorithm is presented in Algorithm 1 in which we compute λ_t that approximates well λ^* and ultimately λ_t^{on} as shown in (15) of Theorem 3.5. We emphasize that the λ_t, $t \in [n]$ computed in Algorithm 1 are obtained by efficiently solving the offline optimization problem (3) given the probabilities p_i up to time $i = t$. Clearly, λ_t returned by Algorithm 1 are not necessarily the same as the optimal λ_t^{on} (see Lemma 3.3). In particular, compared to constraint (10) where the exponents λ_t^{on}, $t \in [n]$ are different for each time step t, constraint (13) has the same exponent λ_t for all time steps $i \in [t]$.

Next, we present the second main result, showing that Algorithm 1 is asymptotically optimal.

Theorem 3.5. *Let q_t^{on}, $t \in [n]$, be the output of Algorithm 1. Let λ^* satisfy (8) such that*

$$Q_t^{\text{off}} = \frac{p_t^{\lambda^*}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda^*}(w)}$$

is the solution to the offline optimization problem (3). If there exists a constant c such that

$$\sum_{i=1}^{t} \text{Var}_{Q_i^{\text{off}}} \left[\log \frac{1}{p_i(w)} \right] \ge ct, \tag{14}$$

then we have:

$$\sum_{t=1}^{n} H\Big(Q_t^{\text{off}}\Big) - \sum_{t=1}^{n} H(q_t^{\text{on}}) \le O(R \log n), \tag{15}$$

$$\sum_{t=1}^{n} \sum_{w \in \mathcal{W}} q_t^{\text{on}} \log p_t(w) + \gamma n \ge O(-R \log n), \tag{16}$$

where $R := R(n) = \max_{t\in[n]} R_t$ *with* R_t *defined in* (12).

Corollary 3.6. *Under condition* (14) *Algorithm 1 satisfies* $\left|\sum_{t=1}^{n} H(Q_t^{\text{on}}) - \sum_{t=1}^{n} H(q_t^{\text{Alg.1}})\right| = \Theta(R(n)\log n)$.

Special Cases. Finally, we study the behavior of R. In general, R can be $O(n)$, however, for many p_t, R reduces to $R = o(n)$. Indeed, consider p_t, $t \in [n]$, that are independently and randomly selected from a finite set of k distributions $\{p^{(1)}, \ldots, p^{(k)}\}$ such that $\Pr(p_t = p^{(i)}) = P_i$, $i \in [k]$ and $\sum_{i=1}^{k} P_i = 1$. While p_t, $t \in [n]$ in practice are not exactly i.i.d., they present some stationarity, which through our analysis for the i.i.d. case, enables Algorithm 1 to converge to a good solution. We start with the following lemma.

Lemma 3.7. *Let* p_t *be independently selected according to* $\Pr(p_t = p^{(i)}) = P_i$, $i \in [k]$ *and* $\hat{P}_i = |\{t : p_t = p^{(i)}\}|/n$, $i \in [k]$ *be the empirical distribution of* p_t *for* $t \in [n]$. *Then with probability at least* $1-\delta$, *the total variation between* $\hat{P}_i$, $i \in [k]$ *and* P_i, $i \in [k]$, *i.e.,* $\frac{1}{2}\sum_{i\in[k]} |\hat{P}_i - P_i|$, *is at most* $\sqrt{\frac{(k+1)\log 2 + \log(1/\delta)}{2n}}$ *for any* k *and* n.

Proof. Note that

$$\frac{\sum_{i\in[k]} |\hat{P}_i - P_i|}{2} = \sup_{A\subseteq[k]} \Big(\sum_{i\in A} \hat{P}_i - \sum_{i\in A} P_i\Big).$$

Since p_t, $t \in [n]$ are i.i.d. generated, by Hoeffding's inequality we have

$$\Pr\left(\Big|\sum_{i\in A} \hat{P}_i - \sum_{i\in A} P_i\Big| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2)$$

for any $A \subseteq [k]$ and $\epsilon > 0$. By the union bound, we have

$$\Pr\left(\sup_{A\subseteq[k]} \Big|\sum_{i\in A} \hat{P}_i - \sum_{i\in A} P_i\Big| \geq \epsilon\right) \leq 2^{k+1}\exp(-2n\epsilon^2).$$

Let $2^{k+1}\exp(-2n\epsilon^2) = \delta$. We have Lemma 3.7. □

We now present our last main result, showing that for p_t, $t \in [n]$, we have $R = o(n)$.

Theorem 3.8. *Let* p_t *be independently selected according to* $\Pr(p_t = p^{(i)}) = P_i$. *Then* $R(n) \leq O\Big(\sqrt{n\log\frac{1}{\delta}}\Big)$ *with probability at least* $1-\delta$.

Proof. W.l.o.g., assume that nP_i, $i \in [k]$ is an integer. Let $\bar{\lambda}$ satisfy

$$\sum_{i\in[k]} nP_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\bar{\lambda}}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\bar{\lambda}}} \log p^{(i)}(w) + n\gamma = 0.$$

Note that

$$\sum_{i\in[k]} n\hat{P}_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + n\gamma = 0.$$

By Lemma 3.7, we have $\sum_{i\in[k]} |\hat{P}_i - P_i| \leq 2\sqrt{\frac{(k+1)\log 2 + \log\frac{1}{\delta}}{2n}}$ with probability at least $1-\delta$. Hence,

$$\begin{aligned}
&\sum_{i\in[k]} nP_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + n\gamma \\
&\leq \sum_{i\in[k]} n\hat{P}_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + n\gamma + \sum_{i\in[k]} n|P_i - \hat{P}_i| \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) \\
&\leq \sqrt{2n\Big((k+1)\log 2 + \log\tfrac{1}{\delta}\Big)}\Big(L_{\max}\log^2\tfrac{1}{\rho} - \gamma\Big),
\end{aligned}$$

where the last inequality follows from Proposition A.1 since $g_t(\lambda) \le L_{\max} \max_\lambda g_t'(\lambda) \le L_{\max} \log^2 \frac{1}{\rho}$.

On the other hand, from Lemma 3.7, we have $\sum_{i\in[k]} |t_i - tP_i| \le \sqrt{2t((k+1)\log 2 + \log \frac{1}{\delta})}$, leading to

$$
\begin{aligned}
&\Big| \sum_{i\in[k]} t_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + t\gamma \Big| \\
&\le \Big| \sum_{i\in[k]} tP_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + t\gamma \Big| + \sum_{i\in[k]} |t_i - tP_i| \Big| \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) \Big| \\
&\le \sqrt{\frac{2t^2((k+1)\log 2 + \log \frac{1}{\delta})}{n}} \Big(L_{\max} \log^2 \tfrac{1}{\rho} - \gamma \Big) + \sqrt{2n\Big((k+1)\log 2 + \log \tfrac{1}{\delta}\Big)} \Big(L_{\max} \log^2 \tfrac{1}{\rho} - \gamma \Big) \\
&\le O\Big(\sqrt{n \log \tfrac{1}{\delta}} \Big)
\end{aligned}
$$

where the last lines follow from Proposition A.1 since $p_t(w) > \rho$. □

4 Experimental Results

We provide a brief empirical validation of our theoretical predictions on the ELI5 dataset [5], using *Meta-Llama-3-8B-Instruct*. Our goal is not to optimize downstream performance, but to verify that the entropy ordering implied by the theory is observed in practice. We compare the offline optimal policy, the proposed online algorithm, and a standard heuristic baseline with fixed temperature $T = 0.7$. We set $\gamma = 0.9$ and apply a probability cutoff of 10^{-4} to obtain truncated distributions.

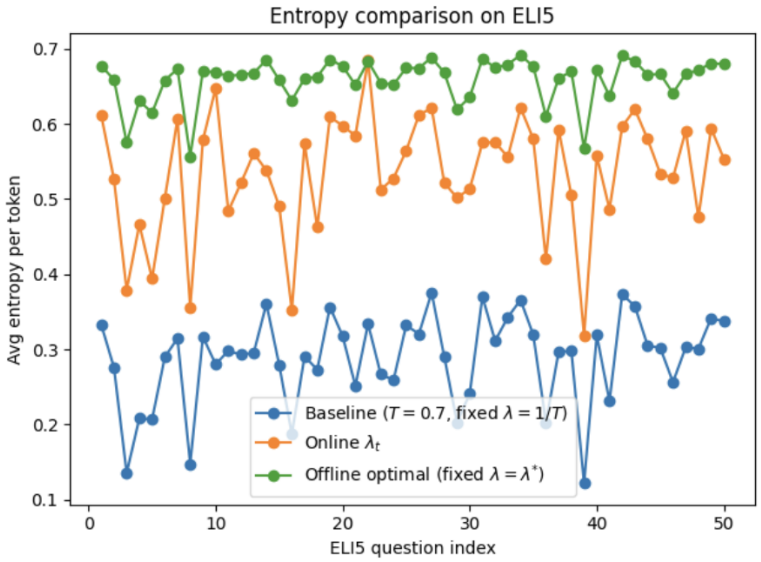

Figure 1: The offline optimum (green) achieves the highest entropy across most questions; the online algorithm (orange) tracks it closely with small gap; the fixed-temperature heuristic (blue) consistently produces lower entropy. The ordering offline ≥ online ≥ heuristic is predicted by Theorems 3.2 and 3.5.

Figure 1 reports the average per-token entropy across sampled instances. The offline optimal achieves the highest entropy, as predicted by Theorem 3.2, while the online algorithm consistently exhibits higher entropy than the heuristic baseline, in agreement with the theoretical guarantees.

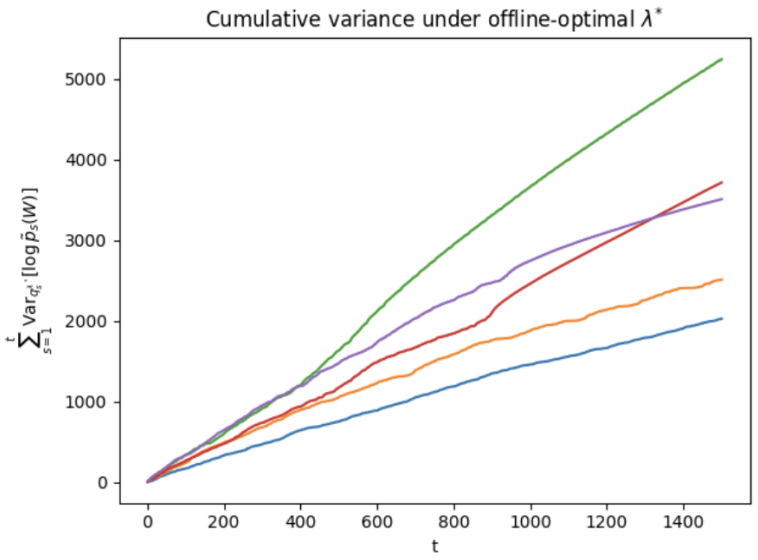

Figure 2: Scaling of the variance $\sum_{i=1}^{t} \mathrm{Var}_{q_i^{\lambda^*}}[\log \tilde{p}_i(W)]$ with respect to t.

Figure 2 shows that the cumulative variance grows approximately linearly with t, providing empirical support for the assumptions underlying Theorem 3.5. Overall, the results are consistent with our theoretical analysis. The approximate linearity of the cumulative variance in Figure 2 provides direct empirical support for the variance condition (14) of Theorem 3.5: condition (14) asserts a linear lower bound, and the empirical curves are well-approximated by a line, so c in (14) can be taken as roughly the slope of these curves.

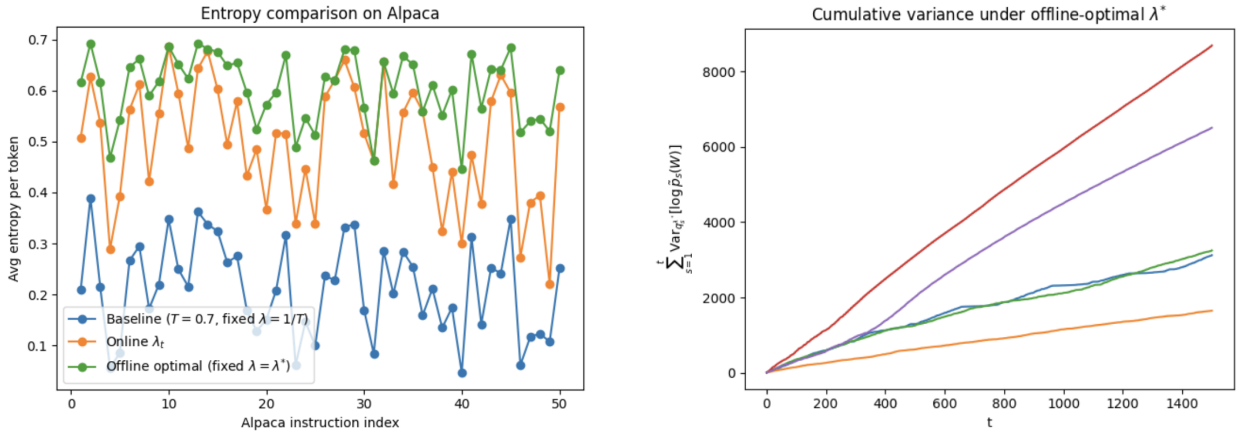

Figure 3: Alpaca. (a) [Left figure] Comparison of per-token entropy for heuristic generation with $T = 0.7$, the proposed online algorithm, and the offline optimal policy. (b) [Right figure] Scaling of the variance $\sum_{i=1}^{t} \mathrm{Var}_{q_i^{\lambda^*}}[\log \tilde{p}_{M,i}(W)]$ with respect to t for 5 sampled generations.

Experiments for Alpaca dataset. Applying the same experimental setup to the Alpaca dataset yields qualitatively similar results. In particular, in the fixed-distribution setting the online algorithm maintains higher entropy than the heuristic baseline, λ_t converges to λ^*, and the cumulative variance exhibits the predicted linear scaling, confirming the generality of our findings. Under independent generations, we observe analogous behavior: the online algorithm continues to attain higher entropy on average, while

maintaining nonnegative plausibility scores as measured by $\log_2 P(W^n) + \gamma n$.

5 Conclusion

We formulated non-repetitive sequence generation as a constrained entropy maximization problem and showed that its offline solution is exactly temperature sampling at a fixed optimal temperature, providing a principled justification for this widely used heuristic. Extending to the online setting, where token distributions are revealed sequentially under a global cross-entropy constraint, we characterized the optimal online solution and designed the Entropic Online Sampling (EOS) algorithm. We proved that EOS incurs entropy loss $O(R \log n)$ relative to the offline optimum with a matching $\Omega(R)$ lower bound, and that $R = o(n)$ under mild stationarity, so EOS is asymptotically optimal. These results give the first rigorous treatment of online constrained entropy maximization for decoding.

A Proof of the Lower Bound (Theorem 3.4)

We first provide some facts about the constraint and the objective function in (1) and (2) that will be repeatedly used in the proofs.

Proposition A.1. *For $p_t(w) > \rho$, let*

$$f_t(\lambda) = -\sum_{w\in\mathcal{W}} \frac{p_t^\lambda(w)}{\sum_{w\in\mathcal{W}} p_t^\lambda(w)} \log \frac{p_t^\lambda(w)}{\sum_{w\in\mathcal{W}} p_t^\lambda(w)}, \tag{17}$$

$$g_t(\lambda) = \sum_{w\in\mathcal{W}} \frac{p_t^\lambda(w)}{\sum_{w\in\mathcal{W}} p_t^\lambda(w)} \log p_t(w) + \gamma. \tag{18}$$

Then we have $0 \le g_t'(\lambda) \le \log^2(\frac{1}{\rho})$ and $-L_{\max}\log^2(\frac{1}{\rho}) \le f_t'(\lambda) \le 0$.

Proof. It can be verified that

$$g_t'(\lambda) = \mathrm{Var}_{\frac{p_t^\lambda(w)}{\sum_{w\in\mathcal{W}} p_t^\lambda(w)}}\left[\log \frac{1}{p_t(w)}\right], \tag{19}$$

$$f_t'(\lambda) = -\lambda g_t'(\lambda). \tag{20}$$

Hence, the proposition follows. □

Lemma A.2. *If $p_t(w) \ge \rho$ for some constant $\rho > 0$, then*

$$\lambda^* \le L_{\max} \triangleq \max\left\{1,\ \frac{\sum_{t=1}^n H(p_t)}{n\gamma + \sum_{t=1}^n \log \max_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}\right\},$$

where λ^* is given in Theorem 3.2. In practice, it is observed that $\sum_{t=1}^n H(p_t) = O(n)$ after filtering out small p_t entries using, e.g., top-k or top-p sampling, and $n\gamma + \sum_{t=1}^n \log \max_{w\in\mathcal{W}} p_t^{\lambda^*}(w) = \Omega(n) > 0$ with γ properly selected. Hence, $L_{\max} \le O(1)$.

Proof. Let

$$f_t(\lambda) = -\sum_{w\in\mathcal{W}} \frac{p_t^\lambda(w)}{\sum_{w\in\mathcal{W}} p_t^\lambda(w)} \log \frac{p_t^\lambda(w)}{\sum_{w\in\mathcal{W}} p_t^\lambda(w)}, \qquad t \in [n]$$

which is the entropy of the distribution

$$q_t(w) = \frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)}.$$

Then

$$\begin{aligned} f_t'(\lambda) &= -\lambda\left(\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)} \log^2 p_t(w) - \Big(\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)} \log p_t(w)\Big)^2 \right) \\ &= -\lambda \operatorname{Var}_{\frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}} \left[\log \frac{1}{p_t(w)}\right] \le 0. \end{aligned} \tag{21}$$

Thus we can write (8) as

$$\begin{aligned} \lambda^*\gamma n &= -\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)} \log p_t^{\lambda^*}(w) = \sum_{t=1}^{n} \Big(f_t(\lambda^*) - \log \sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)\Big) \\ &\le \sum_{t=1}^{n} \Big(f_t(\lambda^*) - \log \max_{w\in\mathcal{W}} p_t^{\lambda^*}(w)\Big) \le \sum_{t=1}^{n} H(p_t) - \lambda^* \sum_{t=1}^{n} \log \max_{w\in\mathcal{W}} p_t^{\lambda^*}(w) \end{aligned}$$

for $\lambda \ge 1$, where the last inequality follows from the facts that $f_t'(\lambda) \le 0$ and that $f_t(1) = H(p_t)$. Hence, we have

$$\lambda^* \le \max\left\{1,\ \frac{\sum_{t=1}^{n} H(p_t)}{n\gamma + \sum_{t=1}^{n} \log \max_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}\right\}$$

which completes the proof of the lemma. □

Now we are ready to prove Theorem 3.4.

Proof of Theorem 3.4. W.l.o.g., let $\mathcal{W} = [|\mathcal{W}|]$. Let M be the smallest integer such that $-\frac{(M-1)\log(\frac{1-\exp(-\gamma/3)}{M-1})}{M} + \frac{\gamma}{3M} > \frac{4\gamma}{3}$ and $\log M \ge \frac{4\gamma}{3}$. Let us define

$$d_x(w) = \begin{cases} x & w = M \\ \frac{1-x}{M-1} & w \in [M-1] \end{cases}$$

for $\frac{1}{M} \le x \le 1$ and let

$$C(\lambda, x) = -\sum_{w\in\mathcal{W}} \frac{d_x^{\lambda}(w)}{\sum_{w\in\mathcal{W}} d_x^{\lambda}(w)} \log d_x(w)$$

for $\lambda \ge 0$. Then

$$\begin{aligned} C\big(0, \exp(-\tfrac{\gamma}{3})\big) &= -\frac{(M-1)\log(\frac{1-\exp(-\gamma/3)}{M-1})}{M} + \frac{\gamma}{3M} > \frac{4\gamma}{3}, \\ C\big(\infty, \exp(-\tfrac{\gamma}{3})\big) &= -\log \max_{w\in\mathcal{W}} d(w) \le \frac{\gamma}{3} < \frac{2\gamma}{3}. \end{aligned}$$

Hence, there exist non-negative λ_1 such that $C(\lambda_1, \exp(-\gamma/3)) = \frac{4\gamma}{3}$ and non-negative λ_2 such that $C(\lambda_2, \exp(-\gamma/3)) = \frac{2\gamma}{3}$, i.e.,

$$\begin{aligned} -\sum_{w\in\mathcal{W}} \frac{d_{\exp(-\gamma/3)}^{\lambda_1}(w)}{\sum_{w\in\mathcal{W}} d_{\exp(-\gamma/3)}^{\lambda_1}(w)} \log d_{\exp(-\gamma/3)}(w) &= \frac{4\gamma}{3}, \\ -\sum_{w\in\mathcal{W}} \frac{d_{\exp(-\gamma/3)}^{\lambda_2}(w)}{\sum_{w\in\mathcal{W}} d_{\exp(-\gamma/3)}^{\lambda_2}(w)} \log d_{\exp(-\gamma/3)}(w) &= \frac{2\gamma}{3}. \end{aligned} \tag{22}$$

Moreover, note that $C(\lambda, \frac{1}{M}) = \log M \geq \frac{4\gamma}{3}$ and $C(\lambda, 1) = 0 \leq \frac{2\gamma}{3}$ for any $\lambda \geq 0$. Hence, there exist $\frac{1}{M} \leq x_1 \leq 1$ such that $C(\lambda_1, x_1) = \frac{2\gamma}{3}$ and $\frac{1}{M} \leq x_2 \leq 1$ such that $C(\lambda_2, x_2) = \frac{4\gamma}{3}$, i.e.,

$$
\begin{aligned}
-\sum_{w\in\mathcal{W}} \frac{d_{x_1}^{\lambda_1}(w)}{\sum_{w\in\mathcal{W}} d_{x_1}^{\lambda_1}(w)} \log d_{x_1}(w) &= \frac{2\gamma}{3}, \\
-\sum_{w\in\mathcal{W}} \frac{d_{x_2}^{\lambda_2}(w)}{\sum_{w\in\mathcal{W}} d_{x_2}^{\lambda_2}(w)} \log d_{x_2}(w) &= \frac{4\gamma}{3}.
\end{aligned} \tag{23}
$$

Moreover, there exist $\frac{1}{M} \leq x_3, x_4 \leq 1$ such that $C(\lambda_1, x_3) = \gamma$ and $C(\lambda_2, x_4) = \gamma$. Consider the following two sequences of distributions

$$
p_{t,1} = \begin{cases} d_{\exp(-\gamma/3)} & t \in [\frac{3R}{\gamma}] \\ d_{x_1} & t \in \{\frac{3R}{\gamma}+1, \ldots, \frac{6R}{\gamma}\} \\ d_{x_3} & t \in \{\frac{6R}{\gamma}+1, \ldots, n\} \end{cases} \quad p_{t,2} = \begin{cases} d_{\exp(-\gamma/3)} & t \in [\frac{3R}{\gamma}] \\ d_{x_2} & t \in \{\frac{3R}{\gamma}+1, \ldots, \frac{6R}{\gamma}\} \\ d_{x_4} & t \in \{\frac{6R}{\gamma}+1, \ldots, n\}. \end{cases} \tag{24}
$$

According to Theorem 3.2, when $p_t = p_{t,\ell}$, $\ell \in \{1, 2\}$, we have

$$
Q_t^{\text{off}} = \frac{p_t^{\lambda_\ell^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_\ell^*}(w)},
$$

where λ_ℓ^*, $\ell \in \{1, 2\}$, are unique solutions satisfying

$$
-\sum_{t\in[n]} \sum_{w\in\mathcal{W}} \frac{p_{t,\ell}^{\lambda_\ell^*}(w)}{\sum_{w\in\mathcal{W}} p_{t,\ell}^{\lambda_\ell^*}(w)} \log p_{t,\ell}(w) = \gamma n.
$$

Note that from (22), (23), and (24), we have

$$
\begin{aligned}
-\sum_{t\in[n]} \sum_{w\in\mathcal{W}} \frac{p_{t,1}^{\lambda_1}(w)}{\sum_{w\in\mathcal{W}} p_{t,1}^{\lambda_1}(w)} \log p_{t,1}(w) &= \frac{3R}{\gamma} \cdot \frac{4\gamma}{3} + \frac{3R}{\gamma} \cdot \frac{2\gamma}{3} + \Big(n - \frac{6R}{\gamma}\Big)\gamma = n\gamma, \\
-\sum_{t\in[n]} \sum_{w\in\mathcal{W}} \frac{p_{t,2}^{\lambda_2}(w)}{\sum_{w\in\mathcal{W}} p_{t,2}^{\lambda_2}(w)} \log p_{t,2}(w) &= \frac{3R}{\gamma} \cdot \frac{2\gamma}{3} + \frac{3R}{\gamma} \cdot \frac{4\gamma}{3} + \Big(n - \frac{6R}{\gamma}\Big)\gamma = n\gamma
\end{aligned}
$$

and thus $\lambda_1^* = \lambda_1$ and $\lambda_2^* = \lambda_2$. In addition, we have

$$
\Big|\sum_{i=1}^{t} \sum_{w\in\mathcal{W}} \frac{p_{i,\ell}^{\lambda_\ell^*}(w)}{\sum_{w\in\mathcal{W}} p_{i,\ell}^{\lambda_\ell^*}(w)} \log p_{i,\ell}(w)\Big| = \begin{cases} \frac{\gamma}{3}\big(\frac{3R}{\gamma} - |t - \frac{3R}{\gamma}|\big) & t \in [\frac{6R}{\gamma}] \\ 0 & t \in \{\frac{6R}{\gamma}+1, \ldots, n\}. \end{cases}
$$

Therefore, we have

$$
\max_{t\in[n]} \Big|\sum_{i=1}^{t} \sum_{w\in\mathcal{W}} Q_i^{\text{off}}(w) \log p_i(w)\Big| = R
$$

both when $p_t = p_{t,1}$, $t \in [n]$, and $p_t = p_{t,2}$, $t \in [n]$.

In the following, we show that

$$
\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \geq \Omega(R)
$$

either when $p_t = p_{t,1}$, $t \in [n]$ or $p_t = p_{t,2}$, $t \in [n]$. Note that $p_t = d_{\exp(-\gamma/3)}$, $t \in [\frac{3R}{\gamma}]$ both when $p_t = p_{t,1}$, $t \in [n]$ and when $p_t = p_{t,2}$. Hence, the sampling distributions $q_t^{\mathcal{A}}$, $t \in [\frac{3R}{\gamma}]$ when $p_t = p_{t,1}$, $t \in [n]$ are the same as $q_t^{\mathcal{A}}$, $t \in [\frac{3R}{\gamma}]$ when $p_t = p_{t,2}$, $t \in [n]$. Denote $Q_{t,\ell}^{\text{off}}$ as the optimal solution to (3) when $p_t = p_{t,\ell}$, $t \in [n]$, $\ell \in \{1, 2\}$. Then from (22) we have

$$\Big| \sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,1}^{\text{off}}(w) \log p_{t,1}(w) - \sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,2}^{\text{off}}(w) \log p_{t,2}(w) \Big| = 2R,$$

which implies

$$\max \left\{ \begin{array}{l} \Big| \Big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,1}^{\text{off}}(w) \log p_{t,1}(w) + 3R \Big) - \Big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_{t,1}(w) + 3R \Big) \Big|, \\ \Big| \Big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,2}^{\text{off}}(w) \log p_{t,2}(w) + 3R \Big) - \Big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_{t,1}(w) + 3R \Big) \Big| \end{array} \right\}$$

$$\geq \frac{\Big| \big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,1}^{\text{off}}(w) \log p_{t,1}(w) + 3R \big) - \big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,2}^{\text{off}}(w) \log p_{t,2}(w) + 3R \big) \Big|}{2} = R.$$

W.l.o.g. assume that

$$\begin{aligned} &\Big| \Big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,1}^{\text{off}}(w) \log p_{t,1}(w) + 3R \Big) - \Big(\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_{t,1}(w) + 3R \Big) \Big| \\ &= \Big| \sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,1}^{\text{off}}(w) \log p_{t,1}(w) - \sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_{t,1}(w) \Big| \geq R. \end{aligned} \tag{25}$$

From (22), we have

$$\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} Q_{t,1}^{\text{off}}(w) \log p_{t,1}(w) + 3R = R. \tag{26}$$

Define

$$B := \sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_{t,1}(w) + 2R.$$

Then according to (25) and (26), we have either $B \geq R$ or $B \leq -R$. We prove that $\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \geq \Omega(R)$ for the case when $p_t = p_{t,1}$, $t \in [n]$ and $B \geq R$. The proof of $\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \geq \Omega(R)$ for the case when $p_t = p_{t,1}$, $t \in [n]$ and $B \leq -R$ is similar.

Let $S(\{p_i\}_{i=1}^t, x)$ be the optimal value of the following optimization problem

$$\max_{q_i \in \Delta(\mathcal{W}): i \in [t]} \sum_{i=1}^{t} H(q_i) \quad \text{s.t.} \quad \sum_{i=1}^{t} \sum_{w \in \mathcal{W}} q_i(w) \log p_i(w) + \gamma t \geq x. \tag{27}$$

Similar to Theorem 3.2, we have $S(\{p_i\}_{i=1}^t, x) = \sum_{i=1}^{t} H(q_i^*)$ where

$$q_i^*(w) = \frac{p_i^{\lambda_x^*}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda_x^*}(w)}, \qquad i \in [t],$$

and λ_x^* satisfy

$$\sum_{i=1}^{t} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda_x^*}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda_x^*}(w)} \log p_i(w) + \gamma t = x.$$

Since $\sum_{t=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_{t,1}(w) + 3R = R + B$, we have $\sum_{t=3R/\gamma+1}^{n} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_{t,1}(w) + n\gamma - 3R = -R - B$ and thus

$$\sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \le S(\{p_i\}_{i=1}^{3R/\gamma}, R + B) + S(\{p_i\}_{i=3R/\gamma+1}^{n}, -R - B).$$

Let λ_R^*, λ_{R+B}^*, λ_{-R}^*, and λ_{-R-B}^* satisfy

$$\begin{aligned}
\sum_{i=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda_R^*}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda_x^*}(w)} \log p_i(w) + 3R &= R \\
\sum_{i=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda_{R+B}^*}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda_{R+B}^*}(w)} \log p_i(w) + 3R &= R + B \\
\sum_{i=3R/\gamma+1}^{n} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda_{-R}^*}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda_{-R}^*}(w)} \log p_i(w) + n - 3R &= -R \\
\sum_{i=3R/\gamma+1}^{n} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda_{-R-B}^*}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda_x^*}(w)} \log p_i(w) + n - 3R &= -R - B. \qquad (28)
\end{aligned}$$

Note that from (26) and the fact that $\sum_{t=1}^{n} \sum_{w \in \mathcal{W}} Q_t^{\text{off}}(w) \log p_t(w) + \gamma n = 0$, we have $\lambda_R^* = \lambda_{-R}^*$. Then

$$\begin{aligned}
&\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \\
&\ge S(\{p_i\}_{i=1}^{3R/\gamma}, R) + S(\{p_i\}_{i=3R/\gamma+1}^{n}, -R) - S(\{p_i\}_{i=1}^{3R/\gamma}, R + B) - S(\{p_i\}_{i=3R/\gamma+1}^{n}, -R - B) \\
&= \sum_{i=1}^{3R/\gamma} f_t(\lambda_R^*) + \sum_{i=3R/\gamma+1}^{n} f_t(\lambda_{-R}^*) - \sum_{i=1}^{3R/\gamma} f_t(\lambda_{R+B}^*) - \sum_{i=3R/\gamma+1}^{n} f_t(\lambda_{-R-B}^*) \\
&= \sum_{i=1}^{3R/\gamma} \int_{\lambda_{R+B}^*}^{\lambda_R^*} f_t'(\lambda)\, d\lambda + \sum_{i=3R/\gamma+1}^{n} \int_{\lambda_{-R-B}^*}^{\lambda_{-R}^*} f_t'(\lambda)\, d\lambda \\
&\overset{(a)}{=} \sum_{i=1}^{3R/\gamma} \int_{\lambda_R^*}^{\lambda_{R+B}^*} \lambda g_t'(\lambda)\, d\lambda + \sum_{i=3R/\gamma+1}^{n} \int_{\lambda_{-R}^*}^{\lambda_{-R-B}^*} \lambda g_t'(\lambda)\, d\lambda \\
&\overset{(b)}{\ge} \sum_{i=1}^{3R/\gamma} \int_{\lambda_R^*}^{\lambda_{R+B}^*} \lambda g_t'(\lambda)\, d\lambda - \sum_{i=3R/\gamma+1}^{n} \int_{\lambda_{-R-B}^*}^{\lambda_{-R}^*} \lambda_{-R}^* g_t'(\lambda)\, d\lambda \\
&\overset{(c)}{=} \sum_{i=1}^{3R/\gamma} \int_{\lambda_R^*}^{\lambda_{R+B}^*} (\lambda - \lambda_R^*) g_t'(\lambda)\, d\lambda \qquad (29)
\end{aligned}$$

where $f_t(\lambda)$ and $g_t(\lambda)$, $t \in [n]$, are defined in (17). Equality (a) follows from the fact that $f_t'(\lambda) = -\lambda g_t'(\lambda)$. Inequality (b) follows from the fact that $\lambda_{-R}^* \ge \lambda_{-R-B}^*$ since $B \ge R \ge 0$ and $g'(\lambda) \ge 0$. Equality (c)

follows from the fact that $\lambda_R^* = \lambda_{-R}^*$ and the fact

$$\int_{\lambda_R^*}^{\lambda_{R+B}^*} g'(\lambda)\, d\lambda = \int_{\lambda_{-R-B}^*}^{\lambda_{-R}^*} g'(\lambda)\, d\lambda = B$$

according to (28).

In the following, we show that $\sum_{i=1}^{3R/\gamma} \int_{\lambda_R^*}^{\lambda_{R+B}^*} (\lambda - \lambda_R^*) g_t'(\lambda)\, d\lambda \geq \Omega(B) = \Omega(R)$. To this end, we first show that $g'(\lambda)$ is lower bounded by $\Omega(1)$ for $\lambda_R^* \leq \lambda \leq \lambda_{3R/2}^* \leq \lambda_{R+B}^*$, where $\lambda_{3R/2}^*$ satisfies

$$\sum_{i=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda_{3R/2}^*}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda_{3R/2}^*}(w)} \log p_i(w) + 3R = \frac{3R}{2}.$$

Note that $g_t(\lambda)$, $t \in [n]$, is increasing in λ. Hence, we have

$$\frac{3R}{2} \leq -\sum_{i=1}^{3R/\gamma} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)} \log p_i(w) \leq 2R$$

and thus

$$\frac{\gamma}{2} \leq -\sum_{w \in \mathcal{W}} \frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)} \log \frac{1}{p_i(w)} \leq \frac{2\gamma}{3}$$

for $\lambda_R^* \leq \lambda \leq \lambda_{3R/2}^* \leq \lambda_{R+B}^*$. Therefore, we have

$$\frac{\gamma}{2} \leq \mathbb{E}_{\frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)}} \left[\log \frac{1}{p_i(w)} \right] \leq \frac{2\gamma}{3}$$

for $t \in [\frac{3R}{\gamma}]$. Note that

$$g_i'(\lambda) = \mathrm{Var}_{\frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)}} \left[\log \frac{1}{p_i(w)} \right],$$

and that $\log \frac{1}{p_i(w)} = \frac{\gamma}{3}$ for $w = M$ and $\log \frac{1}{p_i(w)} \geq \log(M-1) \geq \frac{4\gamma}{3}$. Therefore, we have that

$$g_t'(\lambda) \geq \min \left\{ \left(\tfrac{\gamma}{3} - \tfrac{2\gamma}{3}\right)^2, \left(\tfrac{4\gamma}{3} - \tfrac{2\gamma}{3}\right)^2 \right\} \geq \frac{\gamma^2}{9}.$$

Finally, note that $g_t'(\lambda) \leq \log^2 \frac{1}{p_t(M)} = \log^2 \frac{1}{1-\exp(-\gamma/3)} \leq \left(\frac{4\gamma}{3} + \log \frac{1}{1-\exp(-\gamma/3)}\right)^2$. Hence,

$$\lambda_{3R/2}^* - \lambda_R^* \geq \frac{R}{2\left(\frac{4\gamma}{3} + \log \frac{1}{1-\exp(-\gamma/3)}\right)^2}. \tag{30}$$

From (29) and (30), we have

$$\sum_{t=1}^{n} H(Q_t^{\mathrm{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \geq \sum_{i=1}^{3R/\gamma} \int_{\lambda_R^*}^{\lambda_{R+B}^*} (\lambda - \lambda_R^*) g_t'(\lambda)\, d\lambda \geq \frac{(\lambda_{3R/2}^* - \lambda_R^*)^2 \frac{\gamma^2}{9}}{2} \geq \Omega(R).$$

The proof is complete. □

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[2] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[4] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *Association for Computational Linguistics*, 2018.

[5] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

[6] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural networks as online learners. *arXiv preprint arXiv:1412.7753*, 2014.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, 2017.

[8] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *International Conference on Learning Representations*, 2020.

[9] Mark Kac. On the notion of recurrence in discrete stochastic processes. *Bulletin of the American Mathematical Society*, 53(10):1002–1010, 1947.

[10] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. A diversity-promoting objective function for neural conversation models. *North American Chapter of the Association for Computational Linguistics*, 2016.

[11] Clara Meister and Ryan Cotterell. If beam search is the answer, what was the question? *Transactions of the Association for Computational Linguistics*, 11:152–171, 2023.

[12] Francesco Orabona. Modern online learning: A short survey. *arXiv preprint arXiv:1912.13213*, 2019.

[13] Yaniv Ovadia, Emily Factor, Yarin Molchanov, Dustin Ritter, Rebecca Bixby, Zachary Nado, David Sculley, Sebastian Nowozin, John Duchi, Zoubin Ghahramani, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

[15] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *International Conference on Learning Representations*, 2020.

[16] Changlong Wu, Ananth Grama, and Wojciech Szpankowski. Online universal learning from information-theoretic perspective. *Foundation and Trends in Communication and Information Theory*, 2026. URL https://www.cs.purdue.edu/homes/spa/papers/fnt25.pdf.