Accelerated Speculative Sampling Based on Tree Monte Carlo

Zhengmian Hu[1] **Heng Huang**[1]

Abstract

Speculative Sampling (SpS) has been introduced to speed up the inference of large language models (LLMs) by generating multiple tokens in a single forward pass under the guidance of a reference model, while preserving the original distribution. We observe that SpS can be derived through maximum coupling on the token distribution. However, we find that this approach is not optimal as it applies maximum coupling incrementally for each new token, rather than seeking a global maximum coupling that yields a faster algorithm, given the tree-space nature of LLM generative distributions. In this paper, we shift our focus from distributions on a token space to those on a tree space. We propose a novel class of Tree Monte Carlo (TMC) methods, demonstrating their unbiasedness and convergence. As a particular instance of TMC, our new algorithm, Accelerated Speculative Sampling (ASpS), outperforms traditional SpS by generating more tokens per step on average, achieving faster inference, while maintaining the original distribution.

1. Introduction

Large language models (LLMs) have shown impressive performance across a wide range of tasks such as text generation (Iyer et al., 2022; Chung et al., 2022), translation (Bojar et al., 2017; Barrault et al., 2019), summarization (Liu & Lapata, 2019), *etc*. The remarkable capabilities of these models can largely be attributed to significant increases in both the size of the models and the volume of data which they're trained on. However, a notable downside of these models is their slow inference speed, which limits their applicability in many scenarios.

This challenge has prompted the exploration of methods to accelerate the inference process without compromising the quality of the outcomes. Speculative Sampling (SpS) (Chen et al., 2023; Leviathan et al., 2023) is one recent approach that aims to address this issue. SpS draws from a more efficient reference model to generate what we call reference tokens, and then uses the slower target model to parallelly compute the probabilities of these tokens. Only a subset of these reference tokens are retained as part of the final sampling output. By employing a proper acceptance strategy, it's possible to keep the output distribution of the larger model intact. Moreover, by feeding batches of reference tokens to the target model instead of processing each token with a single forward pass, speculative sampling could increase parallelism and reduce latency. This process takes inspiration from speculative execution in processors, thus earning its name. One of the key benefits of speculative sampling is its ability to accelerate inference without changes to the model architecture or retraining.

In Section 2, we will review the mathematical structure underlying speculative sampling. We consider a joint distribution formed by the sampling distribution of a reference model and the probability that the target model accepts these reference tokens. A reference token is accepted if and only if it is sampled simultaneously by both the reference and target models in this joint distribution. This mechanism is closely related to the concept of maximum coupling, a specific strategy in Monte Carlo simulations which aims to maximize the probability of two or more stochastic processes meeting at the same point. While commonly used in Markov Chain Monte Carlo (MCMC) simulations to merge chains, maximum coupling can also be applied to transfer samples between two distinct distributions, like how speculative sampling aims to translate sampling results from a reference distribution to a target distribution.

In Section 3, we present our key motivation of this paper: speculative sampling is not always optimal. Although SpS is derived from maximum coupling to ensure highest efficiency, it applies maximum coupling incrementally for each new token. Specifically, when generating more than one reference token, the output space becomes a string space, not a mere token space, but a product space of individual token spaces. The maximum coupling within the string space is not simply the product of maximum couplings in each token space. Therefore, SpS is not optimal when the

[1] Department of Computer Science, University of Maryland, College Park, USA. Correspondence to: Zhengmian Hu <huzhengmian@gmail.com>, Heng Huang <henghuanghh@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

number of generated reference tokens exceeds one. We will illustrate this suboptimality and what the optimal solution looks like with an example. We also highlight that the sampling process must take into account the tree structure of the generative space of LLMs, where each node represents a potential token. This is where the concept of Tree Monte Carlo methods comes into play.

In Appendix A, we introduces Tree Monte Carlo (TMC) as methods for sampling paths within a tree-structured space. We define the “tree-distribution” as the distribution among all possible paths within the tree. The most interesting part of TMC, compared to conventional Monte Carlo methods, is that, the order of path introduce a natural partial order among tree-distributions. This leads to the definition of sub-distribution denoted as $P_1 \preceq P_2$. We prove that sub-samplers, which sample from a sub-distribution, could be composed to recover the original distribution. The introduction of sub-samplers opens up new avenues to ensure that the generative process on tree space adheres to the desired distribution.

As a specific instance of TMC methods, we define Accelerated Speculative Sampling (ASpS) in Appendix B, which essentially modifies the sampling outcome of any given tree sampler to construct a new sub-sampler. We demonstrate that ASpS can correctly converge to the target distribution of the target model, relying on the unbiased nature and convergence properties of the sub-samplers. The inputs for ASpS are identical to those used in SpS, involving reference tokens sampled from a reference model, and the probability distributions of both the target and reference models over these reference tokens. The output space remains the same as in SpS, yet ASpS tends to accept a higher average number of tokens, thus leading to faster inference. In the section Section 6, we show that ASpS accepts the same number of tokens as SpS when the number of reference tokens equals one, and it constantly outperforms SpS by accepting more tokens when the number of reference tokens exceeds one.

Due to space limitation, we only show a digest of Appendices A and B in the main paper, and move the complete theory, derivation and proof to the appendix.

We summarize our contributions as follows:

- We propose TMC, which represents a paradigm shift from distributions on token space to distributions on a tree space. This includes a novel concept of sub-distributions that are unique to the tree structure and the convergence theorem of sub-samplers, which lay the foundational work for theoretical guarantees on the convergence of complex sampling methods over tree space. TMC offers a unique perspective to the field of structured sampling algorithms.
- We introduce ASpS, and provide theoretical guarantees that it maintains fidelity to the original distribution as defined by the target model. ASpS requires the same inputs as SpS without relying on additional information, and it achieves faster inference speeds than SpS.

		Reference		
		a	b	c
Target		0.5	0.3	0.2
a	0.3	0.3	0	0
b	0.4	0.1	0.3	0
c	0.3	0.1	0	0.2

Table 1. Single Token Generation with Speculative Sampling

2. Background

We first examine the process of generating a single new token through speculative sampling. Denote the token space as Σ, the reference model as P_{ref}, and the target model as P. Upon sampling a token x from P_{ref}, speculative sampling generates a new token y with the probability:

$$P_{\text{s}}(y;x) = \begin{cases} \min(1, \frac{P(x)}{P_{\text{ref}}(x)}) & \text{if } y = x, \\ \frac{(1-\frac{P(x)}{P_{\text{ref}}(x)})_+(P(y)-P_{\text{ref}}(y))_+}{\sum_{z\in\Sigma}(P_{\text{ref}}(z)-P(z))_+} & \text{if } y \neq x, \end{cases} \qquad (1)$$

where $(\cdot)_+ = \max(0, \cdot)$. The joint distribution $C(x, y) = P_{\text{ref}}(x)P_{\text{s}}(y; x)$, is just the solution to the maximum coupling problem:

$$\max_C \sum_{x\in\Sigma} C(x,x)$$
$$\text{s.t.} \sum_{y\in\Sigma} C(x,y) = P_{\text{ref}}(x), \sum_{x\in\Sigma} C(x,y) = P(y).$$

Consider a toy example in Table 1 where the vocabulary $\Sigma = \{a, b, c\}$. The exact model probabilities are handpicked for demonstration purposes. The couplings are highlighted with colors indicating the number of tokens in agreement, where light green means one matching token and dark green indicates two. If the reference model samples a ‘b’ or ‘c’, speculative decoding deterministically yields the same result. In the case of an ‘a’ being sampled, there’s a 60% chance for ‘a’, and 20% chances each for ‘b’ and ‘c’. This configuration is already optimal and leaves no room for further refinement.

Speculative sampling (SpS) applies maximum coupling incrementally with each new token. Consider a scenario where the reference model, upon sampling ‘a’, samples an additional token. The joint distribution for SpS is shown in Table 2, where the notation $a\tau$ means that SpS accepts ‘a’ and does not proceed to generate another token.

The resulting joint distribution is stepwise product of two maximum coupling. For instance, the probability

				Reference				
				a			b	c
				0.5			0.3	0.2
				aa	ab	ac		
Accepted				0.2	0.2	0.1		
a	0.3	aa	0.1	0.1	0	0		
		ab	0.1	0	0.1	0		
		ac	0.1	0.02	0.02	0.06		
		$a\tau$	0	0	0	0	0	0
b	0.4			0.04	0.04	0.02	0.3	0
c	0.3			0.04	0.04	0.02	0	0.2

Table 2. An Illustration of Speculative Sampling Coupling

				Reference				
				a			b	c
				0.5			0.3	0.2
				aa	ab	ac		
Accepted				0.2	0.2	0.1		
a	0.3	aa	0.1	0.1	0	0		
		ab	0.1	0	0.1	0		
		ac	0.1	0	0	0.1		
		$a\tau$	0	0	0	0	0	0
b	0.4			0.05	0.05	0	0.3	0
c	0.3			0.05	0.05	0	0	0.2

Table 3. An Illustration of Accelerated Speculative Sampling

on the pairing (ac,b) is given by $0.02 = C(ac, b) = C(a, b)P_{\text{ref}}(c|a) = C(a, b)(P_{\text{ref}}(ac)/P_{\text{ref}}(a)) = 0.1 \times (0.1/0.5)$.

3. Preliminary

We observe that SpS is not always optimal. In Table 2, we notice that probability from (ac,b) can be reallocated to enhance the overall strength of the coupling. An improved solution is presented in Table 3, maintaining the marginal distributions while increasing the average number of accepted reference tokens.

Moving forward, this article aims to expand this improvement into a formal method. To achieve this, it becomes necessary to consider distributions on a tree space, where each node represents a potential token. The sampling process must comply with the inherent hierarchical relationships, which inspires the introduction of Tree Monte Carlo in the following section.

				Reference		
				a		b
				0.4		0.6
				aa	ab	
Accepted				0.2	0.2	
a	0.6	aa	$0.1\overline{3}$	$0.1\overline{3}$	0	
		ab	$0.2\overline{6}$	$0.0\overline{6}$	0.2	
		$a\tau$	0.2	0	0	0.2
b	0.4			0	0	0.4

Table 4. Desired Joint Distribution in a Toy Example

For instance, with a vocabulary of just two tokens, $\Sigma = \{a, b\}$, and a target model with $P(a|a) = 1/3$ and $P(b|a) = 2/3$, we desire a joint distribution such as the one in Table 4. Note that when the reference model generates 'b', sampling cannot produce the sequences 'aa' or 'ab'. The marginal distribution of accepted reference tokens does not fully replicate the target model—$0.1\overline{3} = P_{\text{s}}(aa) \neq P(aa) = 0.2$—yet it preserves the proportional relationship between different branches—$\frac{P_{\text{s}}(aa)}{P_{\text{s}}(ab)} = \frac{P(a|a)}{P(b|a)}$. This relationship motivates us to introduce the concept of sub-distribution, which will be discussed in greater detail in the next section.

4. Tree Monte Carlo (TMC)

We start with a quick comparison to conventional Monte Carlo methodology. Traditionally, Monte Carlo methods involve sampling from a space X, which is inherently equipped with a specific distribution P formalized by the Kolmogorov axioms (Ω, F, P), where $\Omega = X$, F is a σ-algebra, and P is a probability measure. To execute the sampling process, one constructs a sampler function $\text{sampler} : B \to X$, where B is a source of randomness that possesses a probability distribution P_B. This source could be, for instance, a perfectly random binary sequence in $B = \{0, 1\}^{\mathbb{N}}$. The goal is to ensure that the result of the sampling procedure $\text{sampler}(b)$ is distributed according to the probability P when the input b follows the distribution P_B.

In short, conventional Monte Carlo methodology takes a probability distribution P defined over a space X, and the goal is to design a sampler $\text{sampler} : B \to X$ such that $\text{sampler}(b) \sim P$ when $b \sim P_B$.

Loosely speaking, Tree Monte Carlo (TMC) considers a tree distribution (Definition A.3) P on the tree-structured space (Definition A.2) X with S as the space of all paths and it aims to design a tree sampler (Definition A.9) $\text{sampler} : B \to S$ such that $\text{sampler}(b) \sim P'$ conforms to a sub-distribution (Definition A.11) $P' \preceq P$ when $b \sim P_B$.

4.1. Tree Distribution and Sub-Distribution

This section will detail the formalism required to mathematically represent tree-structured spaces, tree distributions, and sub-distribution.

Definition 4.1 (Path and Tree Space)**.** A path σ is a finite sequence, with the empty path $\varepsilon = ()$. We define $\sigma_{i:j}$ as the subsequence when $i \leq j$ or ε otherwise, $\sigma + \sigma'$ as the concatenation of σ and σ', $\sigma + a$ as appending element a to σ, and $\sigma \geq \sigma'$ if $\sigma = \sigma' + \sigma''$ for some σ''. $\sigma > \sigma'$ indicates strict extension (i.e., not equal). The length of a path is denoted by $\mathrm{len}(\sigma)$.

A tree space $X = (S, \Sigma)$ comprises a set of paths $S = \{\sigma = (\sigma_1, \ldots, \sigma_n) \mid \sigma_i \in \Sigma(\sigma_{1:i-1}), \forall i \in [n]\}$, and a function Σ representing the finite set of children for each node $\sigma \in S$.

On a tree space X, we can construct the prefix-induced σ-algebra F: For any path $s \in S$, we consider the set of all paths beginning with s, denoted by $S_{\geq s} = \{t \in S \mid t \geq s\}$. Then, we define F to be the smallest σ-algebra that contains $\{S_{\geq s} | s \in S\}$.

Definition 4.2 (Tree distribution)**.** A tree distribution is a probability distribution on the tree space X formally represented as a triple $(S, F, \widehat{P})$, where $F = \sigma(\{S_{\geq s} | s \in S\})$ and $\widehat{P}$ is the probability measure on the σ-algebra F.

We often work with several derived functions rather than directly using $\widehat{P}$:

Definition 4.3 (Path Probability and Support)**.** For a given tree distribution P, the probability of a path $\sigma \in S$ is given by $P(\sigma) = \widehat{P}(S_{\geq \sigma})$. The support of the distribution, denoted as $\mathrm{supp}(P)$, is the set of paths with non-zero probability, i.e., $\mathrm{supp}(P) = \{s \in S | P(s) > 0\}$.

Definition 4.4 (Child Node Probability)**.** For each path $\sigma \in \mathrm{supp}(P)$ and each possible child $a \in \Sigma(\sigma)$, we define the conditional probability of a given σ as $P(a|\sigma) = P(\sigma + a)/P(\sigma)$. Additionally, we introduce a special symbol τ to represent termination at a node, and define $P(\tau|\sigma) = 1 - \sum_{a \in \Sigma(\sigma)} P(a|\sigma) = \widehat{P}(\{\sigma\})/P(\sigma)$, capturing the probability that the path terminates at σ.

Theorem 4.5. *If two tree distributions P_1, P_2 have the same path probabilities, they are the same distribution. That is,*

$$(\forall \sigma \in S, P_1(\sigma) = P_2(\sigma)) \implies \forall U \in F, \widehat{P_1}(U) = \widehat{P_2}(U)$$

Definition 4.6 (Conditional Distributions on a Prefix)**.** Given a prefix path $s \in \mathrm{supp}(P)$, we can induce a new function $\widehat{P_{|s}}(U) = \widehat{P}(U \cap S_{\geq s})/\widehat{P}(S_{\geq s})$, creating a new probability measure that conditions on having a prefix s. Equivalently, for any $a \in \Sigma(\sigma)$, $P_{|s}(a|\sigma)$ is defined as $P(a|\sigma)$ if $\sigma \geq s$ and as the $\mathbb{I}(a = s_{\mathrm{len}(\sigma)+1})$ if $s > \sigma$.

Definition 4.7 (Tree Sampler)**.** A tree sampler is a function $\mathrm{sampler} : B \to S$, where B is a source of randomness with a given probability distribution P_B, mapping each random input to a path in the tree space S.

Definition 4.8 (Induced Tree Distribution by Sampler)**.** Let $\mathrm{sampler} : B \to S$ be a tree sampler and P_B the distribution over B, for any set $U \in F$, the induced probability P_{s} is defined as:

$$\widehat{P_{\mathrm{s}}}(U) = P_B(\mathrm{sampler}(b) \in U).$$

Now we define the crucial concept of a sub-distribution within a tree space.

Definition 4.9 (Sub-distribution)**.** Given two probability distributions on the same tree space X, P_1 and P_2, we say that P_1 is a sub-distribution of P_2, denoted as $P_1 \preceq P_2$, if the following holds:

$$\begin{gathered} \forall \sigma \in \mathrm{supp}(P_1) \cap \mathrm{supp}(P_2), \exists c \in [0, 1], \\ \forall a \in \Sigma(\sigma), P_1(a|\sigma) = cP_2(a|\sigma). \end{gathered} \tag{2}$$

Note that sub-distribution is a partial order among tree-distributions rather than a total order.

Theorem 4.10. *For the sub-distribution $P_1 \preceq P_2$, $\forall \sigma \in S, P_1(\sigma) \leq P_2(\sigma)$.*

It is important to note that sub-distribution relationship is not equivalent to that $P_1(\sigma) \leq P_2(\sigma)$ for all $\sigma \in S$; the latter is just a necessary condition. A sub-distribution relationship captures a more nuanced relation that involves both the probabilities of prefixes and their child node probabilities.

4.2. Sub-Sampler and Its Convergence

We first introduce the concept of leaf-only tree distribution, which conceptually represents the “target” of a tree Monte Carlo process.

Definition 4.11 (Leaf-Only Tree Distribution)**.** A tree distribution is called leaf-only if, for all paths $\sigma \in \mathrm{supp}(P)$ with $\Sigma(\sigma) \neq \phi$, we have that $P(\tau|\sigma) = 0$. This means that the distribution assigns non-zero probability only to the leaf nodes of the tree.

The leaf-only distribution is the maximal element in the ordering of sub-distributions on a tree:

Theorem 4.12. *If $P_1 \preceq P_2$ and P_1 is a leaf-only tree distribution, then P_2 is the same as P_1.*

The intuition is that a leaf-only distribution cannot be further “refined”, and all probabilities are already allocated to the leaves. Therefore it could serve as an endpoint for sequences of tree Monte Carlo procedure.

Now we define the prefix tree sampler, which generates sample that contains a specific prefix.

Definition 4.13 (Prefix Tree Sampler)**.** A prefix tree sampler is a function $\mathrm{sampler} : B \times S \to S$ that takes a random input $b \sim P_B$ and a prefix path $\sigma_{\mathrm{prefix}} \in S$ as inputs, guaranteeing that $\forall b, \sigma_{\mathrm{prefix}}, \mathrm{sampler}(b, \sigma_{\mathrm{prefix}}) \geq \sigma_{\mathrm{prefix}}$.

The induced distribution is expressed as:

$$\widehat{P_{\mathrm{s}}}(U; \sigma_{\mathrm{prefix}}) = P_B(\mathrm{sampler}(b, \sigma_{\mathrm{prefix}}) \in U).$$

Definition 4.14 (Degenerate Prefix Tree Sampler)**.** A prefix tree sampler is called degenerate, if

$$\exists \sigma_{\mathrm{prefix}}, \Sigma(\sigma_{\mathrm{prefix}}) \neq \emptyset \wedge P_{\mathrm{s}}(\tau | \sigma_{\mathrm{prefix}}; \sigma_{\mathrm{prefix}}) = 1.$$

Definition 4.15 (Composition of Prefix Tree Samplers)**.** Considering two prefix tree samplers, $\mathrm{sampler}_{P_\mathrm{s}}$ and $\mathrm{sampler}_{Q_\mathrm{s}}$, we define their composition by:

$$\begin{aligned}&\mathrm{sampler}_{P_\mathrm{s}\circ Q_\mathrm{s}}(b, \sigma_\mathrm{prefix})\\ =&\,\mathrm{sampler}_{P_\mathrm{s}}(\mathrm{split}_1(b), \mathrm{sampler}_{Q_\mathrm{s}}(\mathrm{split}_2(b), \sigma_\mathrm{prefix})),\end{aligned}$$

where b is assumed to be splittable into two independent sources of randomness $\mathrm{split}_1(b)$ and $\mathrm{split}_2(b)$.

We denoted the n times self-composition as P_s^n.

The special case when $n = 0$ corresponds to a sampler that does not extend the path at all:

$$\mathrm{sampler}_{P_\mathrm{s}^0}(b, \sigma_\mathrm{prefix}) = \sigma_\mathrm{prefix}.$$

This composition process is akin to forming Markov chains on the space of tree paths S, where each composition step can be viewed as a transition according to a Markov kernel defined by the prefix tree samplers. Specifically, the composite distribution for a set U is given by:

$$(\widehat{P_\mathrm{s}\circ Q_\mathrm{s}})(U; \sigma_\mathrm{prefix}) = \sum_{\sigma'\in S} \widehat{P_\mathrm{s}}(U; \sigma')\widehat{Q_\mathrm{s}}(\{\sigma'\}; \sigma_\mathrm{prefix}).$$

Definition 4.16 (Prefix Sub-Sampler)**.** A prefix sub-sampler is a type of prefix tree sampler $P_\mathrm{sub}(\cdot; \sigma_\mathrm{prefix})$ that yields a sub-distribution of the original distribution when given any prefix σ_prefix, provided that $\sigma_\mathrm{prefix} \in \mathrm{supp}(P)$. Specifically,

$$\forall \sigma_\mathrm{prefix} \in \mathrm{supp}(P),\ P_\mathrm{sub}(\cdot; \sigma_\mathrm{prefix}) \preceq P_{|\sigma_\mathrm{prefix}}.$$

Theorem 4.17 (Prefix Sub-Samplers is Closed under Composition)**.** *Given a leaf-only tree distribution P and two prefix sub-samplers P_sub and Q_sub, both P_sub^0 and $P_\mathrm{sub} \circ Q_\mathrm{sub}$ are also prefix sub-samplers of P.*

The above theorem highlights a distinct characteristic differentiating Tree Monte Carlo (TMC) from the traditional Markov chain Monte Carlo (MCMC). While traditional MCMC methods also concern the composition of Markov kernels, TMC is unique in that the composition of sub-samplers maintains the order structure between tree distributions. Such order structure is derived from hierarchical nature of tree, and is absent in traditional Monte Carlo.

We can also establish monotone convergence under this composition.

Theorem 4.18 (Monotonicity under Composition of Prefix Sub-Samplers)**.** *Given a tree distribution P and two prefix sub-samplers P_sub and Q_sub, for all paths $\sigma_\mathrm{prefix}, \sigma \in S$, the composition of the two sub-samplers satisfies:*

$$(P_\mathrm{sub} \circ Q_\mathrm{sub})(\sigma; \sigma_\mathrm{prefix}) \geq Q_\mathrm{sub}(\sigma; \sigma_\mathrm{prefix}).$$

It is important to note that a stronger claim would not generally hold. We may wish for a stronger sense of monotonicity, suggesting that Q_sub is a sub-distribution of the composition, stated as

$$\forall \sigma_\mathrm{prefix} \in S, (P_\mathrm{sub} \circ Q_\mathrm{sub})(\cdot; \sigma_\mathrm{prefix}) \succeq Q_\mathrm{sub}(\cdot; \sigma_\mathrm{prefix}),$$

where the inequality is interpreted as sub-distributions. However, such a claim is not universally valid.

Finally, we prove the remarkable convergence property that sub-sampler exhibits.

Theorem 4.19 (Convergence of Non-Degenerate Prefix Sub--Samplers to a Leaf-Only Distribution)**.** *Given a leaf-only tree distribution P, a non-degenerate prefix sub-sampler P_sub, iteratively applied to itself, converges to P in point-wise distance, more formally:*

$$\forall \sigma_\mathrm{p} \in \mathrm{supp}(P), \forall \sigma \in S, \lim_{n\to\infty} P_\mathrm{sub}^n(\sigma; \sigma_\mathrm{p}) = P_{|\sigma_\mathrm{p}}(\sigma).$$

Apart from pointwise convergence for path probability, we can also establish stronger convergence in terms of total variation distance with better uniformity.

Theorem 4.20 (Uniform Convergence of Non-Degenerate Prefix Sub-Samplers to a Leaf-Only Distribution)**.** *Given a leaf-only tree distribution P, a non-degenerate prefix sub-sampler P_sub, iteratively applied to itself, converges to P in total variation distance, more formally:*

$$\forall \sigma_\mathrm{p} \in \mathrm{supp}(P), \lim_{n\to\infty} \mathrm{TV}(\widehat{P_\mathrm{sub}^n}(\cdot; \sigma_\mathrm{p}), \widehat{P_{|\sigma_\mathrm{p}}}) = 0,$$

where the total variation distance is defined as

$$\mathrm{TV}(\widehat{P_\mathrm{sub}^n}(\cdot; \sigma_\mathrm{p}), \widehat{P_{|\sigma_\mathrm{p}}}) = \max_{U\in F}(\widehat{P_\mathrm{sub}^n}(U; \sigma_\mathrm{p}) - \widehat{P_{|\sigma_\mathrm{p}}}(U)).$$

Furthermore, the total variation distance is monotonically decreasing with respect to n.

5. Accelerated Speculative Sampling (ASpS)

After establishing the convergence properties of sub-samplers, the next challenge lies in how to construct such sub-samplers. It is essential to align their design with the specific structure of the problem at hand. In the following section, we will establish SpS and ASpS as examples of sub-samplers.

5.1. Construct Sub-Sampler from a Reference Sampler

Definition 5.1 (Pseudo Leaf-Only Distribution)**.** A tree distribution P is called pseudo leaf-only if $\forall \sigma \in \mathrm{supp}(P), P(\tau|\sigma) \in \{0, 1\}$.

One example of pseudo leaf-only distributions can be obtained by sampling with a reference model up to a predefined depth.

In this section, we address the following task: given a leaf-only target distribution P and a pseudo leaf-only reference distribution P_{ref},how can we construct a sub-distribution $P_{\mathrm{s}} \preceq P$ for speculative sampling?

An additional requirement for speculative sampling is that sampling process must respect the inherent tree structure. For example, if a reference model P_{ref} samples a reference path $\sigma = abc$, under this guidance, it would not be possible to generate the sequence ba as a result. The reason is that feeding the reference path into the target model only yields distributions $P(\cdot|a), P(\cdot|ab), P(\cdot|abc)$, without any knowledge of $P(\cdot|b)$. This observation leads us to define the structure of neighborhoods within the tree space.

Definition 5.2 (Common Prefix)**.** We define operation $\mathrm{cp} : S \times S \to S$ as the common prefix of two paths.

Definition 5.3 (Neighbor Set)**.** Given a sequence σ, we define following sets representing different notions of "neighborhood" in the tree space:

$$S_{\lessapprox\sigma} = \{s \in S |\, \mathrm{len}(\mathrm{cp}(s, \sigma)) \geq \mathrm{len}(s) - 1, \mathrm{len}(s) \leq \mathrm{len}(\sigma)\}$$
$$S_{\approx\sigma} = \{s \in S |\, \mathrm{len}(\mathrm{cp}(s, \sigma)) \geq \mathrm{len}(s) - 1\}$$

Definition 5.4 (Neighbor Distribution)**.** A tree distribution P is called neighbor distribution if the support of $\widehat{P}$ is restricted to a neighbor set. For example, $\widehat{P_{N\lessapprox}}(\{s\}; \sigma) \neq 0$ implies $s \in S_{\lessapprox\sigma}$, and $\widehat{P_{N\approx}}(\{s\}; \sigma) \neq 0$ implies $s \in S_{\approx\sigma}$.

Speculative Sampling (SpS) and Accelerated Speculative Sampling (ASpS) are both neighbor distributions. They can be defined recursively:

$$\widehat{P^*_{N\lessapprox}}(\{\varepsilon\};\varepsilon) = 1,$$
$$m^{\mathrm{SpS}}_{\mathrm{ref}}(a;\sigma) = \widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma) P_{\mathrm{ref}}(a|\sigma)$$
$$m^{\mathrm{ASpS}}_{\mathrm{ref}}(a;\sigma) = P_{\mathrm{ref}}(a|\sigma)$$
$$m^*(a;\sigma) = \widehat{P^*_{N\lessapprox}}(\{\sigma\};\sigma) P(a|\sigma)$$
$$\widehat{P^*_{N\lessapprox}}(\{\sigma + a\};\sigma + a) = \frac{\min(m^*_{\mathrm{ref}}(a;\sigma), m^*(a;\sigma))}{P_{\mathrm{ref}}(a|\sigma)}$$
$$r^*(a;\sigma) = \frac{(m^*_{\mathrm{ref}}(a;\sigma) - m^*(a;\sigma))_+}{\sum_{z\in\Sigma(\sigma)}(m^*_{\mathrm{ref}}(z;\sigma) - m^*(z;\sigma))_+}$$
$$\widehat{P^*_{N\lessapprox}}(\{\sigma + b\};\sigma + a) = \frac{r^*(a;\sigma)(m^*(b;\sigma) - m^*_{\mathrm{ref}}(b;\sigma))_+}{P_{\mathrm{ref}}(a|\sigma)}$$
$$\widehat{P^*_{N\lessapprox}}(\{\sigma\};\sigma + a) = 0$$
$$\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{s\};\sigma + a) = \widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{s\};\sigma)$$
$$\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{s\};\sigma + a) = \frac{r^{\mathrm{ASpS}}(a;\sigma)}{P_{\mathrm{ref}}(a|\sigma)} \widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{s\};\sigma)$$
$$\widehat{P^*_{N\approx}}(\{\sigma + a\};\sigma) = P(a|\sigma) \widehat{P^*_{N\lessapprox}}(\{\sigma\};\sigma)$$
$$\widehat{P^*_{N\approx}}(\{\sigma\};\sigma) = P(\tau|\sigma) \widehat{P^*_{N\lessapprox}}(\{\sigma\};\sigma)$$
$$\widehat{P^*_{N\approx}}(\{s\};\sigma) = \widehat{P^*_{N\lessapprox}}(\{s\};\sigma)$$

where $*$ stands for either SpS or ASpS, and all s in the above satisfies $s \in S_{\lessapprox\sigma} \setminus \{\sigma\}$.

The final sampling distributions for SpS and ASpS are given composing the neighbor distribution and the reference distribution:

$$P^{\mathrm{SpS/ASpS}}_{\mathrm{s}}(\cdot) = \sum_{\sigma\in S} P^{\mathrm{SpS/ASpS}}_{N\approx}(\cdot;\sigma) \widehat{P_{\mathrm{ref}}}(\sigma).$$

For both reference and target distributions, $m_{\mathrm{ref}}(a;\sigma)$ and $m(a;\sigma)$ denote the probability mass that can be used for coupling. Since $m^{\mathrm{ASpS}}_{\mathrm{ref}}(a;\sigma) \geq m^{\mathrm{SpS}}_{\mathrm{ref}}(a;\sigma)$, ASpS achieves a stronger coupling. As an example, the following two theorems illustrate ASpS's superior strength over SpS.

Theorem 5.5. *For every $\sigma \in \mathrm{supp}(P_{\mathrm{ref}})$, it holds that $\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{\sigma\}; \sigma) \geq \widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\}; \sigma)$.*

Theorem 5.6. *If that the expected length of sequences sampled from the reference distribution is finite, i.e., $\mathbb{E}_{s\sim\widehat{P_{\mathrm{ref}}}}[\mathrm{len}(s)] \leq \infty$, we have:*

$$\mathbb{E}_{s\sim\widehat{P^{\mathrm{ASpS}}}}[\mathrm{len}(s)] \geq \mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS}}}}[\mathrm{len}(s)].$$

The following theorem guarantees that both SpS and ASpS can be applied recursively to converge to the target distribution correctly.

Theorem 5.7. *Both SpS and ASpS constitute sub-samplers, i.e., $P^{\mathrm{SpS/ASpS}}_{\mathrm{s}} \preceq P$.*

5.2. Application to LLM

For each LLM, there is a fixed token space Σ, within which there exists a special token known as the end-of-sequence (eos) token, $\mathrm{eos} \in \Sigma$. Use use a predicate, $\mathrm{terminated}(s)$ to represent whether the last token in the sequence s is the eos token. An LLM is a probability distribution as $\mathrm{LLM}(x_{n+1}|x_1, \ldots, x_n)$. A sequence can be continually extended by sampling and appending tokens until the eos token is reached, resulting in a terminated sequence.

A LLM constitutes a tree distribution. Note that all strings forms a tree space:

$$S = (\Sigma \setminus \{\mathrm{eos}\})^* \cup \{s + \mathrm{eos}\,|s \in (\Sigma \setminus \{\mathrm{eos}\})^*\},$$

$$\Sigma(s) = \begin{cases} \phi & \mathrm{terminated}(s), \\ \Sigma & \text{otherwise.} \end{cases}$$

The tree distribution is defined as:

$$P_{\mathrm{LLM}}(a|\sigma) = \begin{cases} 0 & \mathrm{terminated}(\sigma), \\ \mathrm{LLM}(a|\sigma) & \text{otherwise.} \end{cases}$$

Algorithm 1 Accelerated Speculative Sampling (ASpS)

```
**Input:** ref_token : Array[$\Sigma$, len_ref]
   P_ref : Array[Array[float, vocab_size], len_ref]
$\{P_{\text{ref}}(\cdot|\,\text{prefix} + \text{ref\_token}_{1:i}), 0 \le i \le \text{len\_ref} - 1\}$
   P : Array[Array[float, vocab_size], len_ref +1]
$\{P(\cdot|\,\text{prefix} + \text{ref\_token}_{1:i}), 0 \le i \le \text{len\_ref}\}$
len_s ← 0
P_N ← zeros_like(P_ref)
P_N_s_s ← 1
**while** len_s < len_ref **do**
  **if** len_s ≠ 0 **then**
    P_N[len_s, ref_token[len_s]] ← 0
  **end if**
  len_s ← len_s +1
  C1 ← max(0, P_ref[len_s, :] − P_N_s_s × P[len_s, :])
  C2 ← C1[ref_token[len_s]] / sum(C1)
  P_N ← C2 / P_ref[len_s, ref_token[len_s]] × P_N
  P_N[len_s, :] ← C2/P_ref[len_s, ref_token[len_s]]×
      max(0, P_N_s_s × P[len_s, :] − P_ref[len_s, :])
  P_N[len_s, ref_token[len_s]] ←
      min(1, P_N_s_s × P[len_s, ref_token[len_s]]/
      P_ref[len_s, ref_token[len_s]])
  P_N_s_s ← P_N[len_s, ref_token[len_s]]
  assert sum(sum(P_N)) = 1
  $\{P_{N\lessapprox}(\cdot;\sigma)$ is a probability$\}$
**end while**
(gen_len, last_token) ← sample(P_N)
accepted_token ← ref_token[1 : gen_len −1]
accepted_token . append(last_token)
**if** accepted_token = ref_token **then**
  accepted_token . append(sample(P[len_ref +1, :]))
  gen_len ← gen_len +1
**end if**
**Output:** accepted_token : Array[$\Sigma$, gen_len]
```

$$P_{\text{LLM}}(\tau|\sigma) = \begin{cases} 1 & \text{terminated}(\sigma), \\ 0 & \text{otherwise.} \end{cases}$$

With this connection between TMC and LLM, we can apply ASpS to LLM as follows: 1. sample d reference tokens from the reference model that is less computationally expensive. 2. Calculate the neighbor distribution $P_{N\approx}^{\text{ASpS}}(\cdot;\sigma)$. 3. Sample from the neighbor distribution to generate the final output.

6. Implementation

The method defined in the previous section can be translated into pseudo-code in Algorithm 1.

The computational complexity of Algorithm 1 is $\mathcal{O}(\text{len\_ref} \times \text{vocab\_size})$. The primary time consumption of speculative sampling lies in the forward passes of the target model and reference model, which computes the probability for each token. In comparison, the overhead of Algorithm 1 is negligible as it only involves dozen of vector operations.

Table 5. Summary of accepted token numbers per step improvement by ASpS over SpS. n is the number of reference token.

n	SpS	ASpS	Improv. (%)
1	1.284 ± 0.001	1.284 ± 0.001	0.0 ± 0.1
2	1.381 ± 0.002	1.394 ± 0.002	0.9 ± 0.2
3	1.458 ± 0.002	1.485 ± 0.002	1.8 ± 0.2
4	1.454 ± 0.002	1.489 ± 0.002	2.4 ± 0.2
5	1.436 ± 0.002	1.467 ± 0.002	2.1 ± 0.2
6	1.482 ± 0.002	1.507 ± 0.003	1.7 ± 0.2
7	1.477 ± 0.002	1.520 ± 0.003	2.9 ± 0.3
8	1.496 ± 0.003	1.525 ± 0.003	1.9 ± 0.3
9	1.482 ± 0.003	1.532 ± 0.003	3.3 ± 0.3
10	1.468 ± 0.003	1.522 ± 0.003	3.7 ± 0.3

To ensure numerical stability, we implement the algorithm in log-space and handle corner cases such as division by zero appropriately. Our implementations uses Huggingface library (Wolf et al., 2019). It should be noted that this library is not optimized for inference speed. Specifically, the wall-time measurements may not accurately reflect the number of floating-point operations due to memory bandwidth and communication bottleneck.

While assessing the effectiveness of SpS and ASpS, a key metric is the average number of accepted tokens. Importantly, this metric is independent of implementation details, such as the choice of the deep learning library, making it a robust indicator of performance.

In our experiments, we compare the Speculative Sampling (SpS) and Accelerated Speculative Sampling (ASpS) methods using LLaMa-7b model (Touvron et al., 2023) as target model and LLaMa-68m model (Miao et al., 2023) as reference model, on a translation task from the WMT16 dataset (Bojar et al., 2016). The complete experiment results with more tasks and model configurations are moved to Appendix C due to space limitation.

Our experimental results show that when a single reference token is used, the average number of accepted tokens for ASpS is same as that of SpS. This is expected, as the issue with SpS is that it does not apply maximum coupling across the entire space. When there is only one reference token, the token space is effectively the full space, and hence SpS already achieves optimal coupling in this context, leaving no room for ASpS to improve upon.

When the number of reference tokens exceeds one, ASpS consistently accepts more tokens than SpS. This indicates that SpS does not fully leverage the available information and remains sub-optimal. This improved performance underscores the advantage of ASpS.

7. Related Works

7.1. Monte Carlo on Tree Space

The introduction of Tree Monte Carlo (TMC) algorithms presents a unique perspective for sampling in tree-structured space, diverging from both Monte Carlo Tree Search (MCTS) and traditional Markov Chain Monte Carlo (MCMC). Below, we outline the distinctions that separate TMC from these well-established techniques.

7.1.1. Differences from Monte Carlo Tree Search

Monte Carlo Tree Search (MCTS) is a widely recognized algorithm, particularly known for its applications in solving game trees, for instance, in strategic games like Go and Chess. MCTS combines tree search with random sampling and has been fundamental for decision making in complex environments. The essential focus of MCTS is on the winning strategy, wherein the algorithm optimizes the selection of moves that lead to a high probability of winning. It involves a balance of exploration and exploitation, where the search is directed towards more promising sub-trees as more is learned about the game space.

Conversely, Tree Monte Carlo (TMC) algorithms have a different aim: they are designed to sample from a tree distribution, effectively generating a single pathway through a tree space. This process is relevant to structured probabilistic models, where the interest lies in obtaining representative samples from the distribution described by the tree. Unlike MCTS, which specifically concentrates on the decision making aspect in games, TMC is more about statistical inference within hierarchical models.

7.1.2. Differences from Markov Chain Monte Carlo

In traditional Markov Chain Monte Carlo (MCMC), the principle is to construct and iteratively apply a Markov kernel with the goal of obtaining a chain that converges to the desired distribution. Therefore, TMC can be regarded as a special case of MCMC methods. In both cases, the total variation distance to the target distribution monotonically decreases to zero, assuring convergence. However, there are notable differences in the mechanisms and the underpinning mathematical structures that distinguish the two.

TMC introduces a unique form of monotonicity tied to tree structures, which is absent from the mathematical frameworks typically seen in MCMC. TMC also operates distinctly by maintaining prefix sub-samplers which are closed under composition, according to Theorem A.29.

Additionally, TMC does not rely on ergodicity to ensure convergence. This departs from MCMC's dependence on the ergodic condition, which requires that every state is reachable from any other in a finite number of steps. The proof structures of convergence in TMC reflect a different mathematical approach, highlighting unique monotonicity properties associated.

By formalizing the concept of tree distributions and mathematically proving convergence properties specific to sub-sampler, TMC contributes a unique perspective to the field of sampling algorithms, expanding the toolkit available for dealing with structured probabilistic models.

7.2. Accelerating LLM Inference

Speculative sampling (SpS) has emerged as a key advancement for accelerating Large Language Model (LLM) inference without altering the model's output distribution. The foundational principles of SpS, as elucidated by Chen et al. (2023) and Leviathan et al. (2023), operate independently of the reference model's design.

There has been significant exploration into the construction of reference models. For instance, Monea et al. (2023) employs the original model with "look ahead" tokens, while Cai et al. (2023) suggests adding new heads after the last hidden layer of the base model to predict further-ahead tokens. Document retrieval has also been used as a reference model (Yang et al., 2023; He et al., 2023).

One way to extend SpS involves shifting from sequence input to tree input. Standard LLMs take sequences as input, equivalent to a single path through a symbolic tree space. Some studies have modified this approach by using trees as input, allowing multiple branches to be processed in a single forward pass, thus gathering more information to help accelerate decoding (Cai et al., 2023; Spector & Re, 2023; Miao et al., 2023; Yang et al., 2024). This requires altering transformer implementations to replace causal attention with tree attention. Furthermore, speculative sampling itself can be recursively applied, using an additional reference model to expedite the inference of the primary reference model (Spector & Re, 2023).

Beyond SpS, other innovations include sacrificing the unbiased property in exchange for more design flexibility (Kim et al., 2023), or focusing solely on Greedy decoding (Santilli et al., 2023), where generation is modeled as solving a system of equations using methods like Jacobi and Gauss-Seidel iterative techniques for fixed-point finding.

Notably, just like SpS, ASpS is compatible with any reference model, so the improvement is orthogonal to the efforts to enhance the reference model. Moreover, ASpS does not rely on more information than SpS but utilizes the available information more effectively to achieve improvement.

8. Conclusion

We introduced Tree Monte Carlo (TMC), which includes novel sub-distributions unique to tree space and convergence guarantees for sub-samplers. These works provide a theoretical backbone for the complex sampling methods over tree spaces. We also proposed Accelerated Speculative Sampling (ASpS), which enhances SpS without relying on additional information and comes with theoretical guarantee that it retains the target model's distribution. In summary, TMC and ASpS present significant steps forward in efficient sampling methods for tree structures.

Impact Statement

This paper presents work whose goal is to advance the understanding of probabilistic methods on tree space and accelerate the inference speed of large language models. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This work was partially supported by NSF IIS 2347592, 2347604, 2348159, 2348169, DBI 2405416, CCF 2348306, CNS 2347617.

References

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pp. 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4717.

Bojar, O. r., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pp. 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W16/W16-2301`.

Cai, T., Li, Y., Geng, Z., Peng, H., and Dao, T. Medusa: Simple framework for accelerating llm generation with multiple decoding heads, 2023.

Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., and Jumper, J. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

He, Z., Zhong, Z., Cai, T., Lee, J. D., and He, D. Rest: Retrieval-based speculative decoding. *arXiv preprint arXiv:2311.08252*, 2023.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

Kim, S., Mangalam, K., Moon, S., Malik, J., Mahoney, M. W., Gholami, A., and Keutzer, K. Speculative decoding with big little decoder. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Liu, Y. and Lapata, M. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1387.

Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Wong, R. Y. Y., Chen, Z., Arfeen, D., Abhyankar, R., and Jia, Z. Specinfer: Accelerating generative llm serving with speculative inference and token tree verification. *arXiv preprint arXiv:2305.09781*, 2023.

Monea, G., Joulin, A., and Grave, E. Pass: Parallel speculative sampling. *arXiv preprint arXiv:2311.13581*, 2023.

Santilli, A., Severino, S., Postolache, E., Maiorca, V., Mancusi, M., Marin, R., and Rodolà, E. Accelerating transformer inference for translation via parallel decoding. *arXiv preprint arXiv:2305.10427*, 2023.

Spector, B. and Re, C. Accelerating llm inference with staged speculative decoding. *arXiv preprint arXiv:2308.04623*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Yang, N., Ge, T., Wang, L., Jiao, B., Jiang, D., Yang, L., Majumder, R., and Wei, F. Inference with reference: Lossless acceleration of large language models. *arXiv preprint arXiv:2304.04487*, 2023.

Yang, S., Huang, S., Dai, X., and Chen, J. Multi-candidate speculative decoding. *arXiv preprint arXiv:2401.06706*, 2024.

A. Tree Monte Carlo (TMC)

We start with a quick comparison to conventional Monte Carlo methodology. Traditionally, Monte Carlo methods involve sampling from a space X, which is inherently equipped with a specific distribution P formalized by the Kolmogorov axioms (Ω, F, P), where $\Omega = X$, F is a σ-algebra, and P is a probability measure. To execute the sampling process, one constructs a sampler function $\mathrm{sampler} : B \to X$, where B is a source of randomness that possesses a probability distribution P_B. This source could be, for instance, a perfectly random binary sequence in $B = \{0, 1\}^{\mathbb{N}}$. The goal is to ensure that the result of the sampling procedure $\mathrm{sampler}(b)$ is distributed according to the probability P when the input b follows the distribution P_B.

In short, conventional Monte Carlo methodology takes a probability distribution P defined over a space X, and the goal is to design a sampler $\mathrm{sampler} : B \to X$ such that $\mathrm{sampler}(b) \sim P$ when $b \sim P_B$.

Loosely speaking, Tree Monte Carlo (TMC) considers a tree distribution (Definition A.3) P on the tree-structured space (Definition A.2) X with S as the space of all paths and it aims to design a tree sampler (Definition A.9) $\mathrm{sampler} : B \to S$ such that $\mathrm{sampler}(b) \sim P'$ conforms to a sub-distribution (Definition A.11) $P' \preceq P$ when $b \sim P_B$.

A.1. Tree Distribution and Sub-Distribution

This section will detail the formalism required to mathematically represent tree-structured spaces, tree distributions, and sub-distribution.

Definition A.1 (Path)**.** We define a path σ as a finite sequence $(\sigma_1, \ldots, \sigma_n)$. The empty path is denoted by $\varepsilon = ()$. We adopt the following notations for operations on paths:

- **Subsequence**: $\sigma_{i:j}$ refers to the subsequence $(\sigma_i, \sigma_{i+1}, \ldots, \sigma_j)$. If $i > j$, we define it to be the empty path for all $i > j$, i.e., $\sigma_{i:j} = \varepsilon$.
- **Concatenation**: Given two paths $\sigma = (\sigma_1, \ldots, \sigma_n)$ and $\sigma' = (\sigma'_1, \ldots, \sigma'_m)$, we write $\sigma + \sigma'$ to denote the concatenation $(\sigma_1, \ldots, \sigma_n, \sigma'_1, \ldots, \sigma'_m)$.
- **Appending an Element**: For a path $\sigma = (\sigma_1, \ldots, \sigma_n)$ and an element a, we denote $\sigma + a$ as the path extended by a, which is $(\sigma_1, \ldots, \sigma_n, a)$.
- **Order Relation**: A path σ is said to be greater than or equal to a path σ', denoted $\sigma \geq \sigma'$, if there exists a path σ'' such that $\sigma = \sigma' + \sigma''$. The relation $\sigma > \sigma'$ holds if $\sigma \geq \sigma'$ and $\sigma \neq \sigma'$.
- **Length Function**: We define a length function $\mathrm{len} : S \to \mathbb{N}$ assigning the number of elements in the path sequence.

Definition A.2 (Tree Space)**.** A tree space X is defined as a pair (S, Σ), where:

- S is the set of all valid paths, defined by
$$S = \{\sigma = (\sigma_1, \ldots, \sigma_n) \mid \sigma_i \in \Sigma(\sigma_{1:i-1}), \forall i \in [n]\}.$$
- Σ is a function enumerating the children of a node in the tree, where for each $\sigma \in S$, $\Sigma(\sigma)$ is a finite set.

On a tree space X, we can construct the prefix-induced σ-algebra F: For any path $s \in S$, we consider the set of all paths beginning with s, denoted by $S_{\geq s} = \{t \in S \mid t \geq s\}$. Then, we define F to be the smallest σ-algebra that contains $\{S_{\geq s} | s \in S\}$.

Definition A.3 (Tree distribution)**.** A tree distribution is a probability distribution on the tree space X formally represented as a triple $(S, F, \widehat{P})$, where $F = \sigma(\{S_{\geq s} | s \in S\})$ and $\widehat{P}$ is the probability measure on the σ-algebra F.

We often work with several derived functions rather than directly using $\widehat{P}$:

Definition A.4 (Path Probability and Support)**.** For a given tree distribution P, the probability of a path $\sigma \in S$ is given by $P(\sigma) = \widehat{P}(S_{\geq \sigma})$. The support of the distribution, denoted as $\mathrm{supp}(P)$, is the set of paths with non-zero probability, i.e., $\mathrm{supp}(P) = \{s \in S | P(s) > 0\}$.

Definition A.5 (Child Node Probability)**.** For each path $\sigma \in \mathrm{supp}(P)$ and each possible child $a \in \Sigma(\sigma)$, we define the conditional probability of a given σ as $P(a|\sigma) = P(\sigma + a)/P(\sigma)$. Additionally, we introduce a special symbol τ to represent termination at a node, and define $P(\tau|\sigma) = 1 - \sum_{a\in\Sigma(\sigma)} P(a|\sigma) = \widehat{P}(\{\sigma\})/P(\sigma)$, capturing the probability that the path terminates at σ.

Theorem A.6. *If two tree distributions P_1, P_2 have the same path probabilities, they are the same distribution. That is,*

$$(\forall\sigma \in S, P_1(\sigma) = P_2(\sigma)) \implies \forall U \in F, \widehat{P_1}(U) = \widehat{P_2}(U)$$

Proof of Theorem A.6. We observe that $\{S_{\geq s}|s \in S\}$ constitutes a π-system by definition.

We define $\mathcal{L} = \{U \in F|\widehat{P_1}(U) = \widehat{P_2}(U)\}$. It is clear that $\{S_{\geq s}|s \in S\} \subset \mathcal{L} \subset F$. It can be verified that $\mathcal{L}$ is a λ-system (also known as a Dynkin system). Specifically, $\mathcal{L}$ contains S since $\widehat{P_1}(S) = \widehat{P_2}(S) = 1$ by the nature of probability measures. If U is in $\mathcal{L}$, then its complement $S \setminus U$ is also in $\mathcal{L}$ because $\widehat{P_1}(S \setminus U) = 1 - \widehat{P_1}(U) = 1 - \widehat{P_2}(U) = \widehat{P_2}(S \setminus U)$. Furthermore, for a sequence of pairwise disjoint sets $U_1, U_2, \ldots \in \mathcal{L}$, their union $\bigcup_{n=1}^{\infty} U_n$ is also in $\mathcal{L}$ because $\widehat{P_1}(\bigcup_{i=1}^{\infty} U_i) = \sum_{i=1}^{\infty} \widehat{P_1}(U_i) = \sum_{i=1}^{\infty} \widehat{P_2}(U_i) = \widehat{P_2}(\bigcup_{i=1}^{\infty} U_i)$.

By Dynkin's π-λ theorem, $\mathcal{L}$, as a λ-system, must contains the σ-algebra generated by the π-system, that is $F = \sigma(\{S_{\geq s}|s \in S\}) \subset \mathcal{L}$. Therefore we have $F = \mathcal{L}$ and probability measures $\widehat{P_1}$ and $\widehat{P_2}$ are the same for all sets in F. □

As a corollary, one can uniquely specify a tree distribution through path probabilities $P(\sigma)$ alone. Moreover, the tree distribution can be defined solely using child node probabilities $P(a|\sigma)$ for all paths σ in $\mathrm{supp}(P)$, as $P(\sigma) = \prod_{i=1}^{n} P(\sigma_i|\sigma_{1:i-1})$ for all paths with non-zero probability.

Definition A.7 (Conditional Distributions on a Prefix)**.** Given a prefix path $s \in \mathrm{supp}(P)$, we can induce a new function $\widehat{P_{|s}}(U) = \widehat{P}(U \cap S_{\geq s})/\widehat{P}(S_{\geq s})$, creating a new probability measure that conditions on having a prefix s. Equivalently, for any $a \in \Sigma(\sigma)$, $P_{|s}(a|\sigma)$ is defined as $P(a|\sigma)$ if $\sigma \geq s$ and as the $\mathbb{I}(a = s_{\mathrm{len}(\sigma)+1})$ if $s > \sigma$.

Theorem A.8. *If for paths σ_1 and σ_2 in a tree distribution there is a non-zero probability conditioning on σ_1 for path σ_2, i.e., $P_{|\sigma_1}(\sigma_2) \neq 0$, then the paths σ_1 and σ_2 have a definite order relation; explicitly, either $\sigma_1 < \sigma_2$, $\sigma_1 = \sigma_2$, or $\sigma_1 > \sigma_2$.*

Proof of Theorem A.8. By contradiction, assume there is no order relation between σ_1 and σ_2. Then their corresponding sets defined in S do not intersect, more formally, $S_{\geq\sigma_1} \cap S_{\geq\sigma_2} = \emptyset$. This leads to the probability of σ_2 given prefix σ_1 to be zero since $P_{|\sigma_1}(\sigma_2) = \widehat{P_{|\sigma_1}}(S_{\geq\sigma_2}) = \widehat{P}(S_{\geq\sigma_2} \cap S_{\geq\sigma_1})/\widehat{P}(S_{\geq\sigma_1}) = 0$, contradicting the assumption that $P_{|\sigma_1}(\sigma_2) \neq 0$. Thus, σ_1 and σ_2 must have a natural order relation within the tree structure. □

Now we introduce the concept of a tree sampler.

Definition A.9 (Tree Sampler)**.** A tree sampler is a function $\mathrm{sampler} : B \to S$, where B is a source of randomness with a given probability distribution P_B, mapping each random input to a path in the tree space S.

Unlike conventional Monte Carlo samplers, a tree sampler does not necessarily sample paths to their full length every time, as some paths may not extend all the way down to leaf nodes, where $\Sigma(\sigma) = \emptyset$. This characteristic distinguishes tree samplers from standard ones and often necessitates the composition of multiple tree samplers to reach leaf nodes in the tree. We will discuss composition in more detail in the next sub-section.

Upon sampling, the tree sampler induces a distribution over the tree space:

Definition A.10 (Induced Tree Distribution by Sampler)**.** Let $\mathrm{sampler} : B \to S$ be a tree sampler and P_B the distribution over B, for any set $U \in F$, the induced probability P_{s} is defined as:

$$\widehat{P_{\mathrm{s}}}(U) = P_B(\mathrm{sampler}(b) \in U).$$

The path probability P_{s} and the child node probabilities $P_{\mathrm{s}}(a|\sigma)$ induced by the sampler can be computed as follows:

- For any $\sigma \in S$,

$$P_{\mathrm{s}}(\sigma) = P_B(\mathrm{sampler}(b) \in S_{\geq\sigma})$$

.

- For each $\sigma \in \operatorname{supp}(P_\text{s})$, and for each child node $a \in \Sigma(\sigma)$,

$$P_\text{s}(a|\sigma) = \sum_{\sigma' \geq \sigma + a} P_B(\sigma' = \text{sampler}(b))/P_\text{s}(\sigma).$$

- The termination probability at node σ is given by:

$$P_\text{s}(\tau|\sigma) = P_B(\sigma = \text{sampler}(b))/P_\text{s}(\sigma).$$

Now we define the crucial concept of a sub-distribution within a tree space.

Definition A.11 (Sub-distribution)**.** Given two probability distributions on the same tree space X, P_1 and P_2, we say that P_1 is a sub-distribution of P_2, denoted as $P_1 \preceq P_2$, if for every path σ, where child node probabilities $P_1(\cdot|\sigma)$ and $P_1(\cdot|\sigma)$ are both defined, these two probabilities are proportional for all children $a \in \Sigma(\sigma)$. More concisely,

$$\begin{gathered} \forall \sigma \in \operatorname{supp}(P_1) \cap \operatorname{supp}(P_2), \exists c \in [0,1], \\ \forall a \in \Sigma(\sigma), P_1(a|\sigma) = cP_2(a|\sigma). \end{gathered} \tag{3}$$

Note that sub-distribution is a partial order among tree-distributions rather than a total order.

We present the following properties about sub-distributions:

Theorem A.12. *For the sub-distribution $P_1 \preceq P_2$, $\forall \sigma \in S, P_1(\sigma) \leq P_2(\sigma)$.*

Proof of Theorem A.12. The original definition Equation (3) for sub-distribution can be translated into

$$\forall \sigma \in S, P_2(\sigma) \neq 0 \implies P_1(\sigma) = 0 \lor \exists c \in [0,1], \forall a \in \Sigma(\sigma), P_1(a|\sigma) = cP_2(a|\sigma)$$

Once we have established this, we consider two cases separately.

We first consider any σ for which $P_2(\sigma) \neq 0$. If $P_1(\sigma) = 0$, it is immediate that $P_1(\sigma) \leq P_2(\sigma)$. Otherwise, considering $\sigma = (\sigma_1, \ldots, \sigma_n)$, we see that

$$P_1(\sigma) = \prod_{i=1}^{n} P_1(\sigma_i|\sigma_{1:i-1}) \leq \prod_{i=1}^{n} c_i P_2(\sigma_i|\sigma_{1:i-1}) \leq \prod_{i=1}^{n} P_2(\sigma_i|\sigma_{1:i-1}) = P_2(\sigma),$$

where $0 \leq c_i \leq 1$ are the constants dictated by the sub-distribution relationship for each conditional probability along the path. This proves that $P_1(\sigma) \leq P_2(\sigma)$ in the case where $P_2(\sigma) \neq 0$.

Next, for any σ such that $P_2(\sigma) = 0$, we must show that $P_1(\sigma)$ is also 0. By the definition of a path probability as a product of conditional probabilities, there must exist an index i for which $P_2(\sigma_{1:i-1}) \neq 0$ and $P_2(\sigma_i|\sigma_{1:i-1}) = 0$. For this index i, it follows that $P_1(\sigma_{1:i-1}) \leq P_2(\sigma_{1:i-1})$ by the first part of our proof. If $P_1(\sigma_{1:i-1}) = 0$, then $P_1(\sigma_{1:i}) = 0$, and thus, $P_1(\sigma) = 0$. Otherwise, if $P_1(\sigma_{1:i-1}) \neq 0$, then $P_1(\sigma_i|\sigma_{1:i-1})$ must also be 0 since it must be in proportion to $P_2(\sigma_i|\sigma_{1:i-1})$ which is 0. This leads to

$$P_1(\sigma_{1:i}) = P_1(\sigma_i|\sigma_{1:i-1})P_1(\sigma_{1:i-1}) \leq P_2(\sigma_i|\sigma_{1:i-1})P_2(\sigma_{1:i-1}) = P_2(\sigma_{1:i}) = 0.$$

We then have that $P_1(\sigma) = 0$, which also satisfies $P_1(\sigma) \leq P_2(\sigma)$. □

Theorem A.13. $P_1 \preceq P_2 \implies \operatorname{supp}(P_1) \subset \operatorname{supp}(P_2)$.

Theorem A.14. $P_1 \preceq P_2 \implies \forall \sigma \in \operatorname{supp}(P_1) \cap \operatorname{supp}(P_2), P_1(\tau|\sigma) \geq P_2(\tau|\sigma)$.

Proof of Theorem A.13. As a direct result of the order of the path probabilities, $P_2(\sigma) = 0 \implies P_1(\sigma) = 0$, therefore we also derive that the support of P_1 is a subset of the support of P_2, i.e., $\operatorname{supp}(P_1) \subseteq \operatorname{supp}(P_2)$. □

Proof of Theorem A.14. Considering the termination probability, we have:

$$P_1(\tau|\sigma) = 1 - \sum_{a\in\Sigma(\sigma)} P_1(a|\sigma) = 1 - \sum_{a\in\Sigma(\sigma)} cP_2(a|\sigma) \geq 1 - \sum_{a\in\Sigma(\sigma)} P_2(a|\sigma) = P_2(\tau|\sigma),$$

where c is the proportionality constant for the path σ given by the definition of sub-distribution. Since $\sum_{a\in\Sigma(\sigma)} cP_2(a|\sigma) \leq \sum_{a\in\Sigma(\sigma)} P_2(a|\sigma)$ due to $0 \leq c \leq 1$, the inequality holds. □

It is important to note that sub-distribution relationship is not equivalent to that $P_1(\sigma) \leq P_2(\sigma)$ for all $\sigma \in S$; the latter is just a necessary condition. A sub-distribution relationship captures a more nuanced relation that involves both the probabilities of prefixes and their child node probabilities.

We can also consider sub-distributions with respect to a specific prefix path $\sigma_{\text{prefix}} \in S$:

Definition A.15 (Prefix Sub-distribution)**.** For two tree distributions P_1 and P_2 and a prefix path $\sigma_{\text{prefix}} \in S$, we say that P_1 is a sub-distribution of P_2 under prefix σ_{prefix}, if for all paths σ in $S_{\geq\sigma_{\text{prefix}}}$ that are within the support of both P_1 and P_2, there exists a constant $c \in [0, 1]$ such that for all possible extensions $a \in \Sigma(\sigma)$, the conditional probabilities are in proportion:

$$\begin{aligned}\forall\sigma \in S_{\geq\sigma_{\text{prefix}}} \cap \operatorname{supp}(P_1) \cap \operatorname{supp}(P_2), \exists c \in [0, 1],\\ \forall a \in \Sigma(\sigma), P_1(a|\sigma) = cP_2(a|\sigma).\end{aligned}$$

Equivalently, $P_{1|\sigma_{\text{prefix}}} \preceq P_{2|\sigma_{\text{prefix}}}$.

Theorem A.16. *A necessary and sufficient condition for $P_1 \preceq P_2$ is that the following hold:*

$$\begin{aligned}&(\forall\sigma \in S, \forall a, b \in \Sigma(\sigma),\\ &\quad P_1(\sigma + a)P_2(\sigma + b) = P_1(\sigma + b)P_2(\sigma + a))\\ &\wedge(\forall\sigma \in \operatorname{supp}(P_1) \cap \operatorname{supp}(P_2), P_1(\tau|\sigma) \geq P_2(\tau|\sigma))\end{aligned}$$

Proof of Theorem A.16. Firstly, we show the necessity.

We first prove the first clause $(\forall\sigma \in S, \forall a, b \in \Sigma(\sigma), P_1(\sigma + a)P_2(\sigma + b) = P_1(\sigma + b)P_2(\sigma + a))$.

There are a few scenarios to consider:

For every σ in the intersection of the supports of P_1 and P_2, there exists a constant $c \in [0, 1]$ such that for all a in $\Sigma(\sigma)$, $P_1(a|\sigma) = cP_2(a|\sigma)$. Therefore, $P_1(\sigma + a) = P_1(\sigma)P_1(a|\sigma) = P_1(\sigma)cP_2(a|\sigma) = P_1(\sigma)cP_2(\sigma + a)/P_2(\sigma)$. Moreover, $P_1(\sigma + a)P_2(\sigma + b) = P_1(\sigma)cP_2(\sigma + b)P_2(\sigma + a)/P_2(\sigma)$. By a similar argument, $P_1(\sigma + b)P_2(\sigma + a) = P_1(\sigma)cP_2(\sigma + a)P_2(\sigma + b)/P_2(\sigma)$. Due to the common factor, we thus have $P_1(\sigma + a)P_2(\sigma + b) = P_1(\sigma + b)P_2(\sigma + a)$.

For any σ solely in the support of P_2, i.e., $\sigma \in \operatorname{supp}(P_2) \setminus \operatorname{supp}(P_1)$, $P_1(\sigma) = 0$ holds true. Consequently, probabilities of any extensions from such σ are also zero under P_1. This implies $P_1(\sigma + a) = P_1(\sigma + b) = 0$, and consequently, the product form $P_1(\sigma + a)P_2(\sigma + b) = P_1(\sigma + b)P_2(\sigma + a) = 0$ holds.

For any $\sigma \notin \operatorname{supp}(P_2)$, $P_2(\sigma) = 0$, and hence $P_2(\sigma + a) = P_2(\sigma + b) = 0$. This ensures that $P_1(\sigma + a)P_2(\sigma + b) = P_1(\sigma + b)P_2(\sigma + a) = 0$, again holding true.

Next, we prove the second clause $(\forall\sigma \in \operatorname{supp}(P_1) \cap \operatorname{supp}(P_2), P_1(\tau|\sigma) \geq P_2(\tau|\sigma))$. This has already been proven in Theorem A.14.

To prove sufficiency, we need to consider the given conditions in the other direction.

Starting with any $\sigma \in \operatorname{supp}(P_1) \cap \operatorname{supp}(P_2)$ such that there exists an $a \in \Sigma(\sigma)$ with $P_2(\sigma + a) \neq 0$, we define $c = \frac{P_1(\sigma+a)/P_1(\sigma)}{P_2(\sigma+a)/P_2(\sigma)}$. For any $b \in \Sigma(\sigma)$ with $P_2(\sigma + b) \neq 0$, we have that:

$$\frac{P_1(\sigma + b)/P_1(\sigma)}{P_2(\sigma + b)/P_2(\sigma)} = \frac{P_1(\sigma + a)/P_1(\sigma)}{P_2(\sigma + a)/P_2(\sigma)} = c.$$

We need to prove that $c \leq 1$. Due to $P_1(\tau|\sigma) \geq P_2(\tau|\sigma)$, we can see that:

$$P_1(\tau|\sigma) = 1 - \sum_{a\in\Sigma(\sigma)} P_1(a|\sigma) = 1 - \sum_{a\in\Sigma(\sigma)} cP_2(a|\sigma).$$

On the other hand,

$$P_2(\tau|\sigma) = 1 - \sum_{a \in \Sigma(\sigma)} P_2(a|\sigma).$$

Therefore $P_1(\tau|\sigma) \geq P_2(\tau|\sigma)$ leads to the conclusion that $c \leq 1$.

In the case where no children of σ are possible under P_2 (i.e., $P_2(\sigma + a) = 0$ for all $a \in \Sigma(\sigma)$), we set $c = 0$. We must show that $P_1(\sigma + a) = 0$ for all such a. Since $P_2(\tau|\sigma) = 1$ due to there being no children of σ under P_2, and given that $P_1(\tau|\sigma) \geq P_2(\tau|\sigma)$ as stated in the condition, it follows that $P_1(\tau|\sigma) = 1$. This implies that all probabilities of extending σ under P_1 must be 0, meaning $P_1(a|\sigma) = 0$ for all $a \in \Sigma(\sigma)$. Therefore, for all such a, $P_1(\sigma + a) = 0$, satisfying the condition $c = 0$. □

Inspired by Theorem A.12, we define a notion of pointwise distance between two distributions, specifically for a distribution and its sub-distribution, as:

$$\operatorname{dist}_\sigma(P_1, P_2) = P_2(\sigma) - P_1(\sigma).$$

The pointwise distance quantifies how much less likely P_1 is to generate σ compared to P_2. This distance is a measure of the "gap" between them.

Later, we will define the sub-sampler and discuss how multiple sub-samplers can be composed. The most significant theorem we aim to prove is the convergence of sub-samplers, under both pointwise distance and total variation distance.

A.2. Auxiliary Definition and Theorems

Definition A.17. The *cut-off tail probability* $P^{[\mathrm{len}>d]}$ is defined for a tree-distribution P as the probability of all sequences s in the sample space S which have a length exceeding d:

$$P^{[\mathrm{len}>d]} = \widehat{P}(\{s \in S | \operatorname{len}(s) > d\}).$$

Theorem A.18. *For a tree distribution, as* d *approaches infinity, the cut-off tail probability vanishes:*

$$\lim_{d\to\infty} P^{[\mathrm{len}>d]} = 0.$$

Proof of Theorem A.18.

$$\begin{aligned}\lim_{d\to\infty} P^{[\mathrm{len}>d]} &= \lim_{d\to\infty} \widehat{P}(\{s \in S | \operatorname{len}(s) > d\}) \\ &= \widehat{P}(\bigcap_{d=1}^{\infty} \{s \in S | \operatorname{len}(s) > d\}) \\ &= \widehat{P}(\emptyset) \\ &= 0.\end{aligned}$$

□

Theorem A.18 could be useful for computing summation for the whole tree.

Theorem A.19 (Tree Expectation Theorem)**.** *Let* P *be a tree distribution and* $V : S \to \mathbb{R}$ *be a function that satisfies the unbiased condition:*

$$\forall \sigma \in \operatorname{supp}(P), P(\tau|\sigma)V(\sigma) + \sum_{a\in\Sigma(\sigma)} P(a|\sigma)V(\sigma + a) = V(\sigma).$$

Now consider the following scenarios where at least one condition holds:

(a) The tree has a finite depth, that is $\exists d, \forall \sigma \in S, \operatorname{len}(\sigma) \leq d$.

(b) The change in value is bounded and the expected depth is finite. Specifically, $\exists c, \forall \sigma \in S, \forall a \in \Sigma(\sigma), |V(\sigma + a) - V(\sigma)| \leq c$ *and* $\mathbb{E}_{s \in \widehat{P}}[\operatorname{len}(s)] < \infty$.

(c) The function V is bounded across all sequences, i.e., $\exists c, \forall \sigma \in S, |V(\sigma)| \leq c$.

Under these conditions, the expected value of $V(\sigma)$ under the tree distribution $\widehat{P}$ is equal to the value at the root:

$$\mathbb{E}_{\sigma \sim \widehat{P}}[V(\sigma)] = V(\varepsilon).$$

Moreover, the theorem still holds if we change the equality in the unbiased condition to a greater than or equal to (or less than or equal to) relationship, with the result also changing to the corresponding inequality.

Proof of Theorem A.19. Consider $V(\sigma)$ as a stochastic process where $X_t = V(\sigma_{1:\min(\mathrm{len}(\sigma),t)})$. Let the filtration be $F_t = \sigma(\{S_{\geq s} | s \in S, \mathrm{len}(s) \leq t\} \cup \{\{s\} | s \in S, \mathrm{len}(s) \leq t\})$ and stop time as $\tau = \mathrm{len}(\sigma)$. X_t can be observed as a martingale (or submartingale or supermartingale depending on the symbol in the equation) due to the unbiased condition. The conclusion follows directly from applying the optional stopping theorem. □

Theorem A.19 is inspired by the optional stopping theorem found in stochastic process theory and is applied to calculate the expectation over an entire tree.

A.3. Sub-Sampler and Its Convergence

We first introduce the concept of leaf-only tree distribution, which conceptually represents the "target" of a tree Monte Carlo process.

Definition A.20 (Leaf-Only Tree Distribution)**.** A tree distribution is called leaf-only if, for all paths $\sigma \in \mathrm{supp}(P)$ with $\Sigma(\sigma) \neq \phi$, we have that $P(\tau|\sigma) = 0$. This means that the distribution assigns non-zero probability only to the leaf nodes of the tree.

The implication is that non-leaf nodes cannot harbor probability mass. We formalize this intuition with the following theorem:

Theorem A.21. *For a leaf-only tree distribution P, for any paths $\sigma \in S$, if $\widehat{P}(\{\sigma\}) \neq 0$, then we have $\Sigma(\sigma) = \phi$, indicating that σ is a leaf node.*

Proof of Theorem A.21. We will prove the contrapositive: $\forall \sigma \in S$, if $\Sigma(\sigma) \neq \phi$, then $\widehat{P}(\{\sigma\}) = 0$.

Suppose $\Sigma(\sigma) \neq \phi$. We consider two cases:

If $\sigma \in \mathrm{supp}(P)$ for a leaf-only tree distribution, we have $P(\tau|\sigma) = 0$ by definition. Thus, $\widehat{P}(\{\sigma\}) = P(\tau|\sigma)P(\sigma) = 0$.

If $\sigma \notin \mathrm{supp}(P)$, then $\widehat{P}(\{\sigma\}) \leq P(\sigma) = 0$ directly.

In either case, $\widehat{P}(\{\sigma\}) = 0$. □

The leaf-only distribution is the maximal element in the ordering of sub-distributions on a tree:

Theorem A.22. *If $P_1 \preceq P_2$ and P_1 is a leaf-only tree distribution, then P_2 is the same as P_1.*

Proof of Theorem A.22. Since P_1 is a leaf-only tree distribution, all of its probability mass is concentrated on the leaf nodes:

$$\sum_{\sigma \in S} \widehat{P_1}(\{\sigma\}) = \sum_{\sigma \in \{s \in S | \Sigma(s) = \phi\}} \widehat{P_1}(\{\sigma\}) = 1.$$

For every leaf σ where $\Sigma(\sigma) = \phi$, we have $\widehat{P_1}(\{\sigma\}) = P_1(\sigma)$ and $\widehat{P_2}(\{\sigma\}) = P_2(\sigma)$. Due to the sub-distribution relationship, it follows that $\widehat{P_1}(\{\sigma\}) = P_1(\sigma) \leq P_2(\sigma) = \widehat{P_2}(\{\sigma\})$ for every σ.

Considering that the sum of probabilities over all leaf nodes for P_1 equals 1, we can infer that the same sum under distribution P_2 is also 1, thus:

$$1 = \sum_{\sigma \in \{s \in S | \Sigma(s) = \phi\}} \widehat{P_1}(\{\sigma\}) \leq \sum_{\sigma \in \{s \in S | \Sigma(s) = \phi\}} \widehat{P_2}(\{\sigma\}) = 1.$$

The inequality above is thus an equality. This implies that $\widehat{P_1}(\{\sigma\}) = \widehat{P_2}(\{\sigma\})$ for every σ in the support, and P_2 must also be a leaf-only tree distribution. □

The intuition is that a leaf-only distribution cannot be further "refined", and all probabilities are already allocated to the leaves. Therefore it could serve as an endpoint for sequences of tree Monte Carlo procedure.

Now we define the prefix tree sampler, which generates sample conditioned on a prefix.

Definition A.23 (Prefix Tree Sampler)**.** A prefix tree sampler is a function $\mathrm{sampler} : B \times S \to S$ that takes a random input $b \sim P_B$ and a prefix path $\sigma_{\text{prefix}} \in S$ as inputs, guaranteeing that $\forall b, \sigma_{\text{prefix}}, \mathrm{sampler}(b, \sigma_{\text{prefix}}) \geq \sigma_{\text{prefix}}$.

This means that the sampler will output a path in S that contains σ_{prefix} as a prefix. The resulting path has an induced distribution over the tree space, expressed as:

$$\widehat{P_{\mathrm{s}}}(U; \sigma_{\text{prefix}}) = P_B(\mathrm{sampler}(b, \sigma_{\text{prefix}}) \in U).$$

This induced function allows us to calculate path probabilities and child node probabilities:

- **Path Probability**: For a given prefix path σ_{prefix}, we define the path probability of a path σ extending from this prefix as
$$P_{\mathrm{s}}(\sigma; \sigma_{\text{prefix}}) = P_B(\mathrm{sampler}(b, \sigma_{\text{prefix}}) \in S_{\geq \sigma}).$$
- **Child Node Probability**: Additionally, for each σ in the support of $P_{\mathrm{s}}(\cdot; \sigma_{\text{prefix}})$, which is defined as $\{s \in S | P_{\mathrm{s}}(s; \sigma_{\text{prefix}}) > 0\}$, and for all extensions $a \in \Sigma(\sigma)$, the conditional probability is given by
$$P_{\mathrm{s}}(a|\sigma; \sigma_{\text{prefix}}) = P_{\mathrm{s}}(\sigma + a; \sigma_{\text{prefix}}) / P_{\mathrm{s}}(\sigma; \sigma_{\text{prefix}}).$$
- **Termination Probability**: The probability that the path terminates at σ, when σ is prefixed by σ_{prefix}, is
$$\begin{aligned} P_{\mathrm{s}}(\tau|\sigma; \sigma_{\text{prefix}}) &= 1 - \sum_{a \in \Sigma(\sigma)} P_{\mathrm{s}}(a|\sigma; \sigma_{\text{prefix}}) \\ &= \frac{P_B(\sigma = \mathrm{sampler}(b, \sigma_{\text{prefix}}))}{P_{\mathrm{s}}(\sigma; \sigma_{\text{prefix}})}. \end{aligned}$$

Definition A.24 (Degenerate Prefix Tree Sampler)**.** A prefix tree sampler is called degenerate, if for a certain prefix σ_{prefix} where $\Sigma(\sigma_{\text{prefix}}) \neq \emptyset$, always produces a path that terminates at the given prefix, i.e.,

$$\exists \sigma_{\text{prefix}}, \Sigma(\sigma_{\text{prefix}}) \neq \emptyset \wedge P_{\mathrm{s}}(\tau | \sigma_{\text{prefix}}; \sigma_{\text{prefix}}) = 1.$$

In other words, for the given prefix, the degenerate sampler does not extend the path any further; it "stops" the path at the prefix itself.

Definition A.25 (Prefix Tree Sampler Implementing a Tree Distribution)**.** Given a prefix tree sampler $\mathrm{sampler} : B \times S \to S$ that induces a tree distribution $P_{\mathrm{s}}(\cdot; \sigma_{\text{prefix}})$, and a tree distribution P, we say that the prefix tree sampler implements the tree distribution P if

$$\forall \sigma_{\text{prefix}} \in \mathrm{supp}(P),\ P_{|\sigma_{\text{prefix}}}(\cdot) = P_{\mathrm{s}}(\cdot; \sigma_{\text{prefix}}).$$

Our final objective is to construct a prefix tree sampler which implement a leaf-only tree distribution. One way to achieve this is through composition of a lot of tree samplers, each responsible for generating subsequent segments of the paths starting from different prefixes.

Definition A.26 (Composition of Prefix Tree Samplers)**.** Considering two prefix tree samplers, $\mathrm{sampler}_{P_{\mathrm{s}}}$ and $\mathrm{sampler}_{Q_{\mathrm{s}}}$, we define their composition by:

$$\begin{aligned} &\mathrm{sampler}_{P_{\mathrm{s}} \circ Q_{\mathrm{s}}}(b, \sigma_{\text{prefix}}) \\ =& \mathrm{sampler}_{P_{\mathrm{s}}}(\mathrm{split}_1(b), \mathrm{sampler}_{Q_{\mathrm{s}}}(\mathrm{split}_2(b), \sigma_{\text{prefix}})), \end{aligned}$$

where b is assumed to be splittable into two independent sources of randomness $\mathrm{split}_1(b)$ and $\mathrm{split}_2(b)$.

We also define the self-composition of a sampler where we recursively apply the sampler to its own output for n steps, which is denoted as P_{s}^n.

The special case when $n = 0$ corresponds to a sampler that does not extend the path at all:

$$\mathrm{sampler}_{P_{\mathrm{s}}^0}(b, \sigma_{\mathrm{prefix}}) = \sigma_{\mathrm{prefix}}.$$

This composition process is akin to forming Markov chains on the space of tree paths S, where each composition step can be viewed as a transition according to a Markov kernel defined by the prefix tree samplers. Specifically, the composite distribution for a set U is given by:

$$(\widehat{P_{\mathrm{s}} \circ Q_{\mathrm{s}}})(U; \sigma_{\mathrm{prefix}}) = \sum_{\sigma' \in S} \widehat{P_{\mathrm{s}}}(U; \sigma')\widehat{Q_{\mathrm{s}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}}).$$

Path probabilities via composition are then calculated as:

$$(P_{\mathrm{s}} \circ Q_{\mathrm{s}})(\sigma; \sigma_{\mathrm{prefix}}) = \sum_{\sigma' \in S} P_{\mathrm{s}}(\sigma; \sigma')\widehat{Q_{\mathrm{s}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}}).$$

When $n = 0$, the sampler simply returns the input prefix without any further sampling:

$$\widehat{P_{\mathrm{s}}^0}(\{\sigma_{\mathrm{prefix}}\}; \sigma_{\mathrm{prefix}}) = 1,$$
$$P_{\mathrm{s}}^0(\sigma_{\mathrm{prefix}}; \sigma_{\mathrm{prefix}}) = 1,$$
$$P_{\mathrm{s}}^0(\tau|\sigma_{\mathrm{prefix}}; \sigma_{\mathrm{prefix}}) = 1.$$

Theorem A.27 (Invariance of Leaf-Only Distribution Under Composition)**.** *Given two prefix tree samplers* $\mathrm{sampler}_{P_{\mathrm{s}}}$ *and* $\mathrm{sampler}_{Q_{\mathrm{s}}}$*, where* $\mathrm{sampler}_{Q_{\mathrm{s}}}$ *implements the leaf-only tree distribution P, the composed distribution* $(P_{\mathrm{s}} \circ Q_{\mathrm{s}})$ *remains equal to* Q_{s} *for any set U in the σ-algebra F and any* $\sigma_{\mathrm{prefix}} \in \mathrm{supp}(P)$*.*

Proof of Theorem A.27. We first reformulate the theorem we want to prove as

$$\forall U \in F, \forall \sigma_{\mathrm{prefix}} \in \mathrm{supp}(P), (\widehat{P_{\mathrm{s}} \circ Q_{\mathrm{s}}})(U; \sigma_{\mathrm{prefix}}) = \widehat{Q_{\mathrm{s}}}(U; \sigma_{\mathrm{prefix}}).$$

For all prefix paths $\sigma_{\mathrm{prefix}} \in \mathrm{supp}(P)$ and based on the assumption, we have that the induced subset distributions must agree:

$$\widehat{P_{|\sigma_{\mathrm{prefix}}}}(U) = \widehat{Q_{\mathrm{s}}}(U; \sigma_{\mathrm{prefix}}).$$

To prove the composition relation, we need to demonstrate that

$$(\widehat{P_{\mathrm{s}} \circ Q_{\mathrm{s}}})(U; \sigma_{\mathrm{prefix}}) = \sum_{\sigma' \in S} \widehat{P_{\mathrm{s}}}(U; \sigma')\widehat{Q_{\mathrm{s}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}}) = \widehat{P_{|\sigma_{\mathrm{prefix}}}}(U).$$

We observe that for any prefix $\sigma_{\mathrm{prefix}} \in \mathrm{supp}(P)$ and any path $\sigma' \in S$ such that $\widehat{Q_{\mathrm{s}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}}) \neq 0$, we also have $\widehat{P_{|\sigma_{\mathrm{prefix}}}}(\{\sigma'\}) \neq 0$. Therefore, $\widehat{P}(\{\sigma'\}) \neq 0$ and the path σ'. Due to $\widehat{P}(\{\sigma'\}) = P(\sigma')P(\tau|\sigma')$, we also have $P(\tau|\sigma') \neq 0$ and $P(\sigma') \neq 0$. Since P is leaf-only, this implies $P(\tau|\sigma') = 1$. Moreover, the path σ' is in the support of P. This proved the following formula:

$$\forall \sigma_{\mathrm{prefix}} \in \mathrm{supp}(P), \sigma' \in S, \widehat{Q_{\mathrm{s}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}}) \neq 0 \implies P(\tau|\sigma') = 1 \wedge \sigma' \in \mathrm{supp}(P).$$

For any $\sigma' \in \mathrm{supp}(P)$, if $P(\tau|\sigma') = 1$, we have $P_{\mathrm{s}}(\tau|\sigma'; \sigma') = P_{|\sigma'}(\tau|\sigma') = P(\tau|\sigma') = 1$ and $\widehat{P_{\mathrm{s}}}(\{\sigma'\}; \sigma') = P_{\mathrm{s}}(\sigma'; \sigma')P_{\mathrm{s}}(\tau|\sigma'; \sigma') = 1$. Therefore, $\widehat{P_{\mathrm{s}}}(U; \sigma')$ equals 1 if $\sigma' \in U$ and 0 otherwise. This can be represented using the indicator function $\mathbb{I}$ as $\widehat{P_{\mathrm{s}}}(U; \sigma') = \mathbb{I}(\sigma' \in U)$. Specifically, we have that:

$$\forall \sigma' \in \mathrm{supp}(P), P(\tau|\sigma') = 1 \implies \widehat{P_{\mathrm{s}}}(U; \sigma') = \mathbb{I}(\sigma' \in U).$$

Combining the above observations, the composition is computed as:

$$\begin{aligned}\widehat{(P_{\mathrm{s}} \circ Q_{\mathrm{s}})}(U;\sigma_{\text{prefix}}) &= \sum_{\sigma' \in S} \widehat{P_{\mathrm{s}}}(U;\sigma')\widehat{Q_{\mathrm{s}}}(\{\sigma'\};\sigma_{\text{prefix}}) \\ &= \sum_{\sigma' \in S} \mathbb{I}(\sigma' \in U)\widehat{Q_{\mathrm{s}}}(\{\sigma'\};\sigma_{\text{prefix}}) \\ &= \widehat{Q_{\mathrm{s}}}(U;\sigma_{\text{prefix}}) \\ &= \widehat{P_{|\sigma_{\text{prefix}}}}(U).\end{aligned}$$

□

Notably, the above theorem concludes that leaf-only distributions are fixed points under the composition. This inspires us to approximate a desired leaf-only tree distribution by iteratively composing a sequence of prefix samplers. In order to theoretically establish the convergence, we introduce the following definitions.

Definition A.28 (Prefix Sub-Sampler)**.** A prefix sub-sampler is a type of prefix tree sampler $P_{\text{sub}}(\cdot;\sigma_{\text{prefix}})$ that yields a sub-distribution of the original distribution when given any prefix σ_{prefix}, provided that $\sigma_{\text{prefix}} \in \operatorname{supp}(P)$. Specifically,

$$\forall \sigma_{\text{prefix}} \in \operatorname{supp}(P),\ P_{\text{sub}}(\cdot;\sigma_{\text{prefix}}) \preceq P_{|\sigma_{\text{prefix}}}.$$

Theorem A.29 (Prefix Sub-Samplers is Closed under Composition)**.** *Given a leaf-only tree distribution P and two prefix sub-samplers P_{sub} and Q_{sub}, both P_{sub}^0 and $P_{\text{sub}} \circ Q_{\text{sub}}$ are also prefix sub-samplers of P.*

Proof of Theorem A.29. For 0-fold composition, we recall that for $n = 0$, $P_{\text{sub}}^0(\sigma_{\text{prefix}};\sigma_{\text{prefix}}) = 1$ and $P_{\text{sub}}^0(\tau|\sigma_{\text{prefix}};\sigma_{\text{prefix}}) = 1$. We need to show that P_{sub}^0 satisfies the definition of a prefix sub-sampler, i.e., for every $\sigma_{\text{prefix}} \in \operatorname{supp}(P)$, $P_{\text{sub}}^0(\cdot;\sigma_{\text{prefix}}) \preceq P_{|\sigma_{\text{prefix}}}$. Effectively, we must show that for all $\sigma \in S_{\geq \sigma_{\text{prefix}}}$ in the intersection of supports $\operatorname{supp}(P_{\text{sub}}^0(\cdot;\sigma_{\text{prefix}}))$ and $\operatorname{supp}(P)$, there exists a $c \in [0,1]$ such that for all $a \in \Sigma(\sigma)$, $P_{\text{sub}}^0(a|\sigma;\sigma_{\text{prefix}}) = cP(a|\sigma)$. For P_{sub}^0 which does not perform any sampling, we can take $c = 0$.

For one-time composition, we have the composed sampler $(P_{\text{sub}} \circ Q_{\text{sub}})$ where the path probability for the composition is given by the sum of products involving the output from both samplers:

$$(P_{\text{sub}} \circ Q_{\text{sub}})(\sigma;\sigma_{\text{prefix}}) = \sum_{\sigma' \in S} P_{\text{sub}}(\sigma;\sigma')Q_{\text{sub}}(\sigma';\sigma_{\text{prefix}}).$$

To demonstrate that the composition is also a prefix sub-sampler, we have to prove that, for all prefix paths in the support of P, the composed sampler output is a sub-distribution of the original distribution when the prefix is considered, i.e.,

$$\forall \sigma_{\text{prefix}} \in \operatorname{supp}(P),\ (P_{\text{sub}} \circ Q_{\text{sub}})(\cdot;\sigma_{\text{prefix}}) \preceq P_{|\sigma_{\text{prefix}}}.$$

According to Theorem A.16, this is equivalent to

$$\begin{aligned}&\forall \sigma_{\text{prefix}} \in \operatorname{supp}(P), \\ &\quad (\forall \sigma \in S_{\geq \sigma_{\text{prefix}}}, \forall a, b \in \Sigma(\sigma), \\ &\qquad (P_{\text{sub}} \circ Q_{\text{sub}})(\sigma + a;\sigma_{\text{prefix}})P_{|\sigma_{\text{prefix}}}(\sigma + b) \\ &\qquad = (P_{\text{sub}} \circ Q_{\text{sub}})(\sigma + b;\sigma_{\text{prefix}})P_{|\sigma_{\text{prefix}}}(\sigma + a)) \\ &\quad \wedge (\forall \sigma \in S_{\geq \sigma_{\text{prefix}}} \cap \operatorname{supp}((P_{\text{sub}} \circ Q_{\text{sub}})(\cdot;\sigma_{\text{prefix}})) \cap \operatorname{supp}(P_{|\sigma_{\text{prefix}}}), \\ &\qquad (P_{\text{sub}} \circ Q_{\text{sub}})(\tau|\sigma;\sigma_{\text{prefix}}) \geq P_{|\sigma_{\text{prefix}}}(\tau|\sigma)).\end{aligned}$$

We first prove the first half. For any paths $\sigma \geq \sigma_{\text{prefix}}$, and for any children $a, b \in \Sigma(\sigma)$, we will show that that the proportions of extending the prefix by each child are preserved in the composed sub-sampler.

$$(P_{\text{sub}} \circ Q_{\text{sub}})(\sigma + a;\sigma_{\text{prefix}})P_{|\sigma_{\text{prefix}}}(\sigma + b)$$

$$= \sum_{\sigma' \in S} P_{\mathrm{sub}}(\sigma + a; \sigma')\widehat{Q_{\mathrm{sub}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b).$$

According to Theorem A.12 In order for $P_{\mathrm{sub}}(\sigma + a; \sigma')$ to be non-zero, σ' must have a definite order relation with $\sigma + a$, specifically, $\sigma' < \sigma + a$ or $\sigma' \geq \sigma + a$. On the other hand, when $\widehat{Q_{\mathrm{sub}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}})$ is non-zero, σ' must be either equal to or extend σ_{prefix}, that is $\sigma' \geq \sigma_{\mathrm{prefix}}$.

Breaking down the sum, we have:

$$\begin{aligned}
&(P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})(\sigma + a; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) \\
=& \sum_{\sigma_{\mathrm{prefix}} \leq \sigma' < \sigma + a} P_{\mathrm{sub}}(\sigma + a; \sigma')\widehat{Q_{\mathrm{sub}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) \\
&+ \sum_{\sigma' \geq \sigma + a} P_{\mathrm{sub}}(\sigma + a; \sigma')\widehat{Q_{\mathrm{sub}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) \\
=& \sum_{\sigma_{\mathrm{prefix}} \leq \sigma' \leq \sigma} P_{\mathrm{sub}}(\sigma + a; \sigma')\widehat{Q_{\mathrm{sub}}}(\{\sigma'\}; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) \\
&+ Q_{\mathrm{sub}}(\sigma + a; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b).
\end{aligned}$$

In case $\sigma' \in \mathrm{supp}(P)$, we have

$$P_{\mathrm{sub}}(\sigma + a; \sigma')P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) = P_{\mathrm{sub}}(\sigma + a; \sigma')P_{|\sigma'}(\sigma + b)P_{|\sigma_{\mathrm{prefix}}}(\sigma').$$

Due to the fact that P_{sub} is a prefix sub-sampler, we have

$$P_{\mathrm{sub}}(\sigma + a; \sigma')P_{|\sigma'}(\sigma + b) = P_{\mathrm{sub}}(\sigma + b; \sigma')P_{|\sigma'}(\sigma + a).$$

Therefore

$$P_{\mathrm{sub}}(\sigma + a; \sigma')P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) = P_{\mathrm{sub}}(\sigma + b; \sigma')P_{|\sigma_{\mathrm{prefix}}}(\sigma + a).$$

In case $\sigma' \notin \mathrm{supp}(P)$, we have $P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) = P_{|\sigma_{\mathrm{prefix}}}(\sigma + a) = 0$ and $P_{\mathrm{sub}}(\sigma + a; \sigma')P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) = P_{\mathrm{sub}}(\sigma + b; \sigma')P_{|\sigma_{\mathrm{prefix}}}(\sigma + a)$.

Combining either case, we proved that

$$P_{\mathrm{sub}}(\sigma + a; \sigma')P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) = P_{\mathrm{sub}}(\sigma + b; \sigma')P_{|\sigma_{\mathrm{prefix}}}(\sigma + a).$$

Moreover, due to the fact that Q_{sub} is a prefix sub-sampler, we have

$$Q_{\mathrm{sub}}(\sigma + a; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) = Q_{\mathrm{sub}}(\sigma + b; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + a).$$

Thus, combining the factors, we conclude that:

$$(P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})(\sigma + a; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + b) = (P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})(\sigma + b; \sigma_{\mathrm{prefix}})P_{|\sigma_{\mathrm{prefix}}}(\sigma + a).$$

Now for the second part, we consider the termination probabilities for $\sigma \in S_{\geq \sigma_{\mathrm{prefix}}}$ when σ is within the intersection of supports for the composed distribution $(P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})(\cdot; \sigma_{\mathrm{prefix}})$ and $P_{|\sigma_{\mathrm{prefix}}}$. We need to show that the termination probability for the composed distribution is at least as large as for the original prefix-distribution:

$$(P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})(\tau|\sigma; \sigma_{\mathrm{prefix}}) \geq P_{|\sigma_{\mathrm{prefix}}}(\tau|\sigma).$$

Given that P is a leaf-only tree distribution and considering the classification of paths σ, we distinguish two cases:

If $\Sigma(\sigma) \neq \phi$, implying σ is not a leaf, it holds that $P_{|\sigma_{\mathrm{prefix}}}(\tau|\sigma) = 0$ by the definition of a leaf-only distribution. Therefore, it is trivial that $(P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})(\tau|\sigma; \sigma_{\mathrm{prefix}}) \geq P_{|\sigma_{\mathrm{prefix}}}(\tau|\sigma)$.

If $\Sigma(\sigma) = \phi$, implying σ is a leaf, the termination probabilities for both $P_{|\sigma_{\mathrm{prefix}}}$ and $(P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})$ at σ are 1. Hence, we have equality,

$$(P_{\mathrm{sub}} \circ Q_{\mathrm{sub}})(\tau|\sigma; \sigma_{\mathrm{prefix}}) = P_{|\sigma_{\mathrm{prefix}}}(\tau|\sigma) = 1.$$

□

The above theorem highlights a distinct characteristic differentiating Tree Monte Carlo (TMC) from the traditional Markov chain Monte Carlo (MCMC). While traditional MCMC methods also concern the composition of Markov kernels, TMC is unique in that the composition of sub-samplers maintains the order structure between tree distributions. Such order structure is derived from hierarchical nature of tree structure, and is absent in traditional Monte Carlo.

We can also establish monotone convergence under this composition.

Theorem A.30 (Monotonicity under Composition of Prefix Sub-Samplers)**.** *Given a tree distribution P and two prefix sub-samplers P_{sub} and Q_{sub}, for all paths $\sigma_{\text{prefix}}, \sigma \in S$, the composition of the two sub-samplers satisfies:*

$$(P_{\text{sub}} \circ Q_{\text{sub}})(\sigma; \sigma_{\text{prefix}}) \geq Q_{\text{sub}}(\sigma; \sigma_{\text{prefix}}).$$

Proof of Theorem A.30. The path probability of the composition $(P_{\text{sub}} \circ Q_{\text{sub}})$ is given by taking the sum over all paths σ' that are children of σ with respect to the sub-sampler P_{sub}, weighted by the probability that σ' is reached from σ_{prefix} using the sub-sampler Q_{sub}:

$$\begin{aligned}
&(P_{\text{sub}} \circ Q_{\text{sub}})(\sigma; \sigma_{\text{prefix}}) \\
=&\sum_{\sigma' \in S} P_{\text{sub}}(\sigma; \sigma')\widehat{Q_{\text{sub}}}(\{\sigma'\}; \sigma_{\text{prefix}}) \\
\geq &\sum_{\sigma' \in S_{\geq \sigma}} P_{\text{sub}}(\sigma; \sigma')\widehat{Q_{\text{sub}}}(\{\sigma'\}; \sigma_{\text{prefix}}) \\
=&\sum_{\sigma' \in S_{\geq \sigma}} \widehat{Q_{\text{sub}}}(\{\sigma'\}; \sigma_{\text{prefix}}) \\
=&\widehat{Q_{\text{sub}}}(S_{\geq \sigma}; \sigma_{\text{prefix}}) \\
=&Q_{\text{sub}}(\sigma; \sigma_{\text{prefix}})
\end{aligned}$$

□

Essentially, P_{sub} acts as a refining step that extend the path σ_{prefix} in a way that respect target distribution P.

It is important to note that a stronger claim would not generally hold. We may wish for a stronger sense of monotonicity, suggesting that Q_{sub} is a sub-distribution of the composition, stated as

$$\forall \sigma_{\text{prefix}} \in S, (P_{\text{sub}} \circ Q_{\text{sub}})(\cdot; \sigma_{\text{prefix}}) \succeq Q_{\text{sub}}(\cdot; \sigma_{\text{prefix}}),$$

where the inequality is interpreted as sub-distributions. However, such a claim is not universally valid.

Finally, we prove the remarkable convergence property that sub-sampler exhibits.

Theorem A.31 (Convergence of Non-Degenerate Prefix Sub-Samplers to a Leaf-Only Distribution)**.** *Given a leaf-only tree distribution P, a non-degenerate prefix sub-sampler P_{sub}, iteratively applied to itself, converges to P in pointwise distance, more formally:*

$$\forall \sigma_{\text{p}} \in \text{supp}(P), \forall \sigma \in S, \lim_{n \to \infty} P^n_{\text{sub}}(\sigma; \sigma_{\text{p}}) = P_{|\sigma_{\text{p}}}(\sigma).$$

Proof of Theorem A.31. For simple cases, the path probabilities are already aligned with what is required:

- If the path σ is neither a prefix nor an extension of σ_{prefix}, then the probability is zero and remains so. That is $P^n_{\text{sub}}(\sigma; \sigma_{\text{prefix}}) = 0$ and $P_{|\sigma_{\text{prefix}}}(\sigma) = 0$.
- If the path $\sigma \leq \sigma_{\text{prefix}}$, then the probability remains one, $P^n_{\text{sub}}(\sigma; \sigma_{\text{prefix}}) = 1$, $P_{|\sigma_{\text{prefix}}}(\sigma) = 1$, as per definition.

Excluding the simple cases, we assume now that $\sigma > \sigma_{\text{prefix}}$, and denote $\sigma = \sigma_{\text{prefix}} + (a_1, \ldots, a_n)$.

We define vectors b_i and c_i representing the probabilities calculated after i applications of the sub-sampler:

$$b_i = \begin{bmatrix} P^i_{\text{sub}}(\sigma_{\text{prefix}} + a_{1:1}; \sigma_{\text{prefix}}) \\ P^i_{\text{sub}}(\sigma_{\text{prefix}} + a_{1:2}; \sigma_{\text{prefix}}) \\ \vdots \\ P^i_{\text{sub}}(\sigma_{\text{prefix}} + a_{1:n}; \sigma_{\text{prefix}}) \end{bmatrix}, c_i = \begin{bmatrix} \widehat{P^i_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:0}\}; \sigma_{\text{prefix}}) \\ \widehat{P^i_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:1}\}; \sigma_{\text{prefix}}) \\ \vdots \\ \widehat{P^i_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:n-1}\}; \sigma_{\text{prefix}}) \end{bmatrix}. \tag{4}$$

We then characterize the relationship between b_i and c_i. Let $b_{i,j}$ denote the j-th component of b_i and $c_{i,j}$ denote the j-th component of c_i, which are computed as follows:

$$b_{i,j} = P^i_{\text{sub}}(\sigma_{\text{prefix}} + a_{1:j}; \sigma_{\text{prefix}}),$$

$$c_{i,j} = \widehat{P^i_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:j-1}\}; \sigma_{\text{prefix}}).$$

The relationship between each $b_{i,j}$ and $c_{i,j}$ captures the way the path probabilities depends on the the probabilities of intermediate nodes. Specifically, we have

$$\begin{aligned} & b_{i,j} \\ =& P^i_{\text{sub}}(\sigma_{\text{prefix}} + a_{1:j}; \sigma_{\text{prefix}}) \\ =& [P^i_{\text{sub}}(\sigma_{\text{prefix}} + a_{1:j-1}; \sigma_{\text{prefix}}) - \widehat{P^i_{\text{sub}}}(\sigma_{\text{prefix}} + a_{1:j-1}; \sigma_{\text{prefix}})] \\ & \times P(a_j | \sigma_{\text{prefix}} + a_{1:j-1}) \\ =& \begin{cases} (b_{i,j-1} - c_{i,j}) P(a_j | \sigma_{\text{prefix}} + a_{1:j-1}) & j > 1 \\ (1 - c_{i,j}) P(a_1 | \sigma_{\text{prefix}}) & j = 1 \end{cases} \end{aligned}$$

The second equality is based on the fact that P_{sub} is a sub-sampler of P.

Expanding these recursive relationship, we have

$$b_{i,j} = P_{|\sigma_{\text{prefix}}}(\sigma_{\text{prefix}} + a_{1:j}) - \sum_{k=1}^{j} c_{i,k} P_{|\sigma_{\text{prefix}} + a_{1:k-1}}(\sigma_{\text{prefix}} + a_{1:j})$$

In matrix form, we have

$$b_i = \begin{bmatrix} P_{|\sigma_{\text{prefix}}}(\sigma_{\text{prefix}} + a_{1:1}) \\ P_{|\sigma_{\text{prefix}}}(\sigma_{\text{prefix}} + a_{1:2}) \\ \vdots \\ P_{|\sigma_{\text{prefix}}}(\sigma_{\text{prefix}} + a_{1:n}) \end{bmatrix} - A c_i,$$

$$A_{i,j} = \mathbb{I}(i \geq j) P_{|\sigma_{\text{prefix}} + a_{1:j-1}}(\sigma_{\text{prefix}} + a_{1:i}).$$

Then we compute how c_i change over iteration:

$$\begin{aligned} & c_{i+1,j} \\ =& \widehat{P^{i+1}_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:j-1}\}; \sigma_{\text{prefix}}) \\ =& \sum_{\sigma' \in S} \widehat{P_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:j-1}\}; \sigma') \widehat{P_{\text{sub}}}(\{\sigma'\}; \sigma_{\text{prefix}}) \\ =& \sum_{\sigma_{\text{prefix}} \leq \sigma' \leq \sigma_{\text{prefix}} + a_{1:j-1}} \widehat{P_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:j-1}\}; \sigma') \widehat{P_{\text{sub}}}(\{\sigma'\}; \sigma_{\text{prefix}}) \\ =& \sum_{k=1}^{j} \widehat{P_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:j-1}\}; \sigma_{\text{prefix}} + a_{1:k-1}) \widehat{P_{\text{sub}}}(\{\sigma_{\text{prefix}} + a_{1:k-1}\}; \sigma_{\text{prefix}}) \end{aligned}$$

$$=\sum_{k=1}^{j}\widehat{P_{\text{sub}}}(\{\sigma_{\text{prefix}}+a_{1:j-1}\};\sigma_{\text{prefix}}+a_{1:k-1})c_{i,k}$$

In matrix form, we have

$$c_{i+1}=Bc_i,$$

$$B_{i,j}=\mathbb{I}(i\geq j)\widehat{P_{\text{sub}}}(\{\sigma_{\text{prefix}}+a_{1:i-1}\};\sigma_{\text{prefix}}+a_{1:j-1}).$$

The iteration matrix described above is a lower triangular matrix with diagonal elements greater than 0 and less than 1, as P_{sub} is a prefix sub-sampler of P.

Given the lower triangular structure and the bounds on the diagonal, we can conclude that the limit of c_i as i approaches infinity exists and equals zero for each j, that is

$$\lim_{i\to\infty}c_{i,j}=0. \tag{5}$$

With the limits of c_i established as zeroes, the limits of b_i exist and match the distribution $P_{|\sigma_{\text{prefix}}}$ for every extended path from the prefix σ_{prefix}. Hence we have:

$$\lim_{i\to\infty}b_{i,j}=P_{|\sigma_{\text{prefix}}}(\sigma_{\text{prefix}}+a_{1:j})$$

for each segment $a_{1:j}$ of the path extending σ_{prefix}.

Therefore, we have proven

$$\forall\sigma_{\text{prefix}}\in\text{supp}(P),\forall\sigma\in S,\lim_{n\to\infty}P^n_{\text{sub}}(\sigma;\sigma_{\text{prefix}})=P_{|\sigma_{\text{prefix}}}(\sigma).$$

□

Apart from pointwise convergence for path probability, we can also establish stronger convergence in terms of total variation distance with better uniformity.

Theorem A.32 (Uniform Convergence of Non-Degenerate Prefix Sub-Samplers to a Leaf-Only Distribution)**.** *Given a leaf-only tree distribution P, a non-degenerate prefix sub-sampler P_{sub}, iteratively applied to itself, converges to P in total variation distance, more formally:*

$$\forall\sigma_{\text{p}}\in\text{supp}(P),\lim_{n\to\infty}\text{TV}(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{p}}),\widehat{P_{|\sigma_{\text{p}}}})=0,$$

where the total variation distance is defined as

$$\text{TV}(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{p}}),\widehat{P_{|\sigma_{\text{p}}}})=\max_{U\in F}(\widehat{P^n_{\text{sub}}}(U;\sigma_{\text{p}})-\widehat{P_{|\sigma_{\text{p}}}}(U)).$$

Furthermore, the total variation distance is monotonically decreasing with respect to n.

Proof of Theorem A.32. First, we reformulate the total variation distance on the leaves of the tree:

- For all $\sigma\in S$ such that $\Sigma(\sigma)=\phi$ (leaf nodes):

$$\widehat{P^n_{\text{sub}}}(\{\sigma\};\sigma_{\text{prefix}})=P^n_{\text{sub}}(\sigma;\sigma_{\text{prefix}})\leq P_{|\sigma_{\text{prefix}}}(\sigma)=\widehat{P_{|\sigma_{\text{prefix}}}}(\{\sigma\}).$$

- For all $\sigma\in S$ such that $\Sigma(\sigma)\neq\phi$ (non-leaf nodes):

$$\widehat{P^n_{\text{sub}}}(\{\sigma\};\sigma_{\text{prefix}})\geq\widehat{P_{|\sigma_{\text{prefix}}}}(\{\sigma\})=0.$$

Therefore, the total variation distance can be described as follows:

$$\begin{aligned}&\mathrm{TV}(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})\\ =&\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi\}}\\ =&1-\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)=\phi\}}\end{aligned}$$

From this reformulation, we can prove the monotonicity of total variation distance by noticing that for all $\sigma\in\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)=\phi\}$, the probability $\widehat{P_{\text{sub}}^{n}}(\{\sigma\};\sigma_{\text{prefix}})=P_{\text{sub}}^{n}(\sigma;\sigma_{\text{prefix}})$ is monotonically increasing as n increases, due to Theorem A.30. As a result, the total variation distance $\mathrm{TV}(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})$ is monotonically decreasing with respect to n.

To handle the convergence, we can approach the total variation distance by considering it up to a certain length and the tail probability after that length.

Let us define the cut-off total variation distance as:

$$\mathrm{TV}_d(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})=\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\mathrm{len}(s)\leq d\}}.$$

Also, following Definition A.17, let us define the cut-off tail probability as:

$$P_{|\sigma_{\text{prefix}}}^{[\mathrm{len}>d]}=\widehat{P_{|\sigma_{\text{prefix}}}}(\{s\in S|\,\mathrm{len}(s)>d\}).$$

With these definitions, we can bound the total variation distance as follows:

$$\begin{aligned}&\mathrm{TV}_d(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})\\ \leq&\,\mathrm{TV}(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})\\ \leq&\,\mathrm{TV}_d(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})+P_{|\sigma_{\text{prefix}}}^{[\mathrm{len}>d]}.\end{aligned}$$

The left-hand side of the above claim is clear by definition:

$$\begin{aligned}&\max_{U\in F}(\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})-\widehat{P_{|\sigma_{\text{prefix}}}}(U))\\ \geq&(\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})-\widehat{P_{|\sigma_{\text{prefix}}}}(U))|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\mathrm{len}(s)\leq d\}}\\ =&\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\mathrm{len}(s)\leq d\}}\end{aligned}$$

For the right-hand side, we have:

$$\begin{aligned}&\mathrm{TV}(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})\\ =&\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi\}}\\ =&\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\mathrm{len}(s)\leq d\}}\\ &+\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\mathrm{len}(s)>d\}}\\ =&\,\mathrm{TV}_d(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})\\ &+\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\mathrm{len}(s)>d\}}\\ \leq&\,\mathrm{TV}_d(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})+\widehat{P_{\text{sub}}^{n}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|\,\mathrm{len}(s)>d\}}\\ =&\,\mathrm{TV}_d(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})+\sum_{\sigma\in S,\mathrm{len}(\sigma)=d+1}P_{\text{sub}}^{n}(\sigma;\sigma_{\text{prefix}})\\ \leq&\,\mathrm{TV}_d(\widehat{P_{\text{sub}}^{n}}(\cdot;\sigma_{\text{prefix}}),\widehat{P_{|\sigma_{\text{prefix}}}})+\sum_{\sigma\in S,\mathrm{len}(\sigma)=d+1}P_{|\sigma_{\text{prefix}}}(\sigma)\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{TV}_d(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{prefix}}), \widehat{P_{|\sigma_{\text{prefix}}}}) + \widehat{P_{|\sigma_{\text{prefix}}}}(U)|_{U=\{s\in S|\ \text{len}(s)>d\}} \\
&= \mathrm{TV}_d(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{prefix}}), \widehat{P_{|\sigma_{\text{prefix}}}}) + P^{[\text{len}>d]}_{|\sigma_{\text{prefix}}}
\end{aligned}
$$

As shown in Theorem A.18, the tail probability vanishes as the cut-off length goes to infinity:

$$\lim_{d\to\infty} P^{[\text{len}>d]}_{|\sigma_{\text{prefix}}} = 0$$

We also have that as $n \to \infty$, the total variation distance over paths that are up to length d converges to zero. For a given length d, consider the set of all paths that extend σ_{prefix} but are not themselves leaves, and are of length $\leq d$:

$$\forall \sigma \in \{s \in S | s \geq \sigma_{\text{prefix}}, \Sigma(s) \neq \phi, \text{len}(s) \leq d\}.$$

From our previous result regarding the convergence of the path probability, specifically Equations (4) and (5), we know that as $n \to \infty$, for each σ in the set:

$$\lim_{n\to\infty} \widehat{P^n_{\text{sub}}}(\{\sigma\};\sigma_{\text{prefix}}) = 0.$$

Using this, we can show that the cut-off total variation distance converges to zero:

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{TV}_d(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{prefix}}), \widehat{P_{|\sigma_{\text{prefix}}}}) \\
&= \lim_{n\to\infty} \widehat{P^n_{\text{sub}}}(U;\sigma_{\text{prefix}})|_{U=\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\text{len}(s)\leq d\}} \\
&= \lim_{n\to\infty} \widehat{P^n_{\text{sub}}}(\{s \in S | s \geq \sigma_{\text{prefix}}, \Sigma(s) \neq \phi, \text{len}(s) \leq d\};\sigma_{\text{prefix}}) \\
&= \lim_{n\to\infty} \sum_{\sigma\in\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\text{len}(s)\leq d\}} \widehat{P^n_{\text{sub}}}(\{\sigma\};\sigma_{\text{prefix}}) \\
&= \sum_{\sigma\in\{s\in S|s\geq\sigma_{\text{prefix}},\Sigma(s)\neq\phi,\text{len}(s)\leq d\}} \lim_{n\to\infty} \widehat{P^n_{\text{sub}}}(\{\sigma\};\sigma_{\text{prefix}}) \\
&=0
\end{aligned}
$$

The exchange of limits and summation is allowed because the set over which we are summing is finite for a fixed length d.

Finally, we prove that the total variation distance converges to zero as $n \to \infty$. For every positive integer d, the total variation distance is bounded by the cut-off total variation distance plus the tail probability:

$$
\begin{aligned}
\forall d \geq 1,\ &\lim_{n\to\infty} \mathrm{TV}(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{prefix}}), \widehat{P_{|\sigma_{\text{prefix}}}}) \\
&\leq \lim_{n\to\infty} (\mathrm{TV}_d(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{prefix}}), \widehat{P_{|\sigma_{\text{prefix}}}}) + P^{[\text{len}>d]}_{|\sigma_{\text{prefix}}}) = P^{[\text{len}>d]}_{|\sigma_{\text{prefix}}}.
\end{aligned}
$$

Taking limits as $d \to \infty$ and using the earlier observation that the tail probability vanishes, we get:

$$\lim_{n\to\infty} \mathrm{TV}(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{prefix}}), \widehat{P_{|\sigma_{\text{prefix}}}}) \leq \inf_{d\geq 1} P^{[\text{len}>d]}_{|\sigma_{\text{prefix}}} \leq \lim_{d\to\infty} P^{[\text{len}>d]}_{|\sigma_{\text{prefix}}} = 0.$$

On the other hand, by definition, the total variation distance must be non-negative, so we have:

$$\lim_{n\to\infty} \mathrm{TV}(\widehat{P^n_{\text{sub}}}(\cdot;\sigma_{\text{prefix}}), \widehat{P_{|\sigma_{\text{prefix}}}}) = 0.$$

This proves the uniform convergence in total variation distance of the non-degenerate prefix sub-sampler to the leaf-only distribution P. □

B. Accelerated Speculative Sampling (ASpS)

After establishing the convergence properties of sub-samplers, the next challenge lies in how to construct such sub-samplers. It is essential to align their design with the specific structure of the problem at hand. In the following section, we will establish SpS and ASpS as examples of sub-samplers.

B.1. Construct Sub-Sampler from a Reference Sampler

Definition B.1 (Pseudo Leaf-Only Distribution)**.** A tree distribution P is called pseudo leaf-only if $\forall \sigma \in \mathrm{supp}(P), P(\tau|\sigma) \in \{0, 1\}$. A pseudo leaf σ is one where $P_{\mathrm{ref}}(\tau|\sigma) = 1$.

One example of leaf-only distributions can be obtained by sampling with a reference model over a predefined number of steps.

In this section, we address the following task: given a leaf-only target distribution P and a pseudo leaf-only reference distribution P_{ref},[1] how can we output a sub-distribution $P_{\mathrm{s}} \preceq P$?

An additional requirement for speculative sampling is that sampling process must respect the inherent tree structure. For example, if a reference model P_{ref} samples a reference path $\sigma = abc$, under this guidance, it would not be possible to generate the sequence ba as a result. The reason is that feeding the reference path into the target model only yields distributions $P(\cdot|a), P(\cdot|ab), P(\cdot|abc)$, without any knowledge of $P(\cdot|b)$. This observation leads us to define the structure of neighborhoods within the sampling space.

Definition B.2 (Common Prefix)**.** The common prefix operation $\mathrm{cp} : S \times S \to S$ is defined for two paths $\sigma_1, \sigma_2 \in S$ as follows: If $\sigma_1 = \sigma + (a_1, \ldots, a_{n_1})$, $\sigma_2 = \sigma + (b_1, \ldots, b_{n_2})$, and $a_1 \neq b_1$, their largest common prefix is σ. If $\sigma_1 \leq \sigma_2$ or $\sigma_1 \geq \sigma_2$, the largest common prefix is the shorter of the two paths.

Definition B.3 (Neighbor Set)**.** Given a sequence σ, we define several sets representing different notions of "neighborhood" in the tree space:

$$S_{<\sigma} = \{s \in S | s < \sigma\}$$

$$S_{\leq\sigma} = \{s \in S | s \leq \sigma\}$$

$$S_{\simeq\sigma} = \{s \in S | \operatorname{len}(\mathrm{cp}(s,\sigma)) \geq \operatorname{len}(s) - 1, \operatorname{len}(s) = \operatorname{len}(\sigma)\}$$

$$S_{\gtrsim\sigma} = \{s \in S | \operatorname{len}(\mathrm{cp}(s,\sigma)) \geq \operatorname{len}(s) - 1, \operatorname{len}(s) = \operatorname{len}(\sigma) + 1\}$$

$$S_{\lesssim\sigma} = \{s \in S | \operatorname{len}(\mathrm{cp}(s,\sigma)) \geq \operatorname{len}(s) - 1, \operatorname{len}(s) \leq \operatorname{len}(\sigma) - 1\}$$

For convenience, $S_{\approx\sigma} = S_{\simeq\sigma} \cup S_{\gtrsim\sigma} \cup S_{\lesssim\sigma}$ and $S_{\lessapprox\sigma} = S_{\simeq\sigma} \cup S_{\lesssim\sigma}$ are also defined. The length condition $\operatorname{len}(\mathrm{cp}(s,\sigma)) \geq \operatorname{len}(s) - 1$ implies that the sequence s matches with σ up to at least its penultimate element.

Definition B.4 (Neighbor Distribution)**.** A tree distribution P is called neighbor distribution if its support ($\mathrm{supp}(\widehat{P})$ instead of $\mathrm{supp}(P)$) is restricted to a neighbor set. For example, $\widehat{P_{N\lessapprox}}(\{s\};\sigma) \neq 0$ implies $s \in S_{\lessapprox\sigma}$, and $\widehat{P_{N\approx}}(\{s\};\sigma) \neq 0$ implies $s \in S_{\approx\sigma}$.

Speculative Sampling (SpS) and Accelerated Speculative Sampling (ASpS) are both neighbor distributions. They can be defined recursively:

$$\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\varepsilon\};\varepsilon) = 1,$$

$$m^{\mathrm{SpS}}_{\mathrm{ref}}(a;\sigma) = \widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma) P_{\mathrm{ref}}(a|\sigma)$$

$$m^{\mathrm{ASpS}}_{\mathrm{ref}}(a;\sigma) = P_{\mathrm{ref}}(a|\sigma)$$

$$m^{\mathrm{SpS/ASpS}}(a;\sigma) = \widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma) P(a|\sigma)$$

$$\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma + a\};\sigma + a) = \frac{\min(m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(a;\sigma), m^{\mathrm{SpS/ASpS}}(a;\sigma))}{P_{\mathrm{ref}}(a|\sigma)}$$

$$r^{\mathrm{SpS/ASpS}}(a;\sigma) = \frac{(m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(a;\sigma) - m^{\mathrm{SpS/ASpS}}(a;\sigma))_+}{\sum_{z \in \Sigma(\sigma)} (m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(z;\sigma) - m^{\mathrm{SpS/ASpS}}(z;\sigma))_+}$$

[1] The choice of pseudo leaf-only reference distribution is made for simplicity, but rest assured, the derived Adaptive Speculative Sampling (ASpS) is still applicable to non-pseudo leaf-only distributions, albeit potentially with sub-optimal sampling efficiency.

$$\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma+b\};\sigma+a)=\frac{r^{\mathrm{SpS/ASpS}}(a;\sigma)}{P_{\mathrm{ref}}(a|\sigma)}(m^{\mathrm{SpS/ASpS}}(b;\sigma)-m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(b;\sigma))_+$$

$$\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma+a)=0$$

$$\forall s\in S_{\lessapprox\sigma}\setminus\{\sigma\},\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{s\};\sigma+a)=\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{s\};\sigma)$$

$$\forall s\in S_{\lessapprox\sigma}\setminus\{\sigma\},\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{s\};\sigma+a)=\frac{r^{\mathrm{ASpS}}(a;\sigma)}{P_{\mathrm{ref}}(a|\sigma)}\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{s\};\sigma)$$

$$\widehat{P^{\mathrm{SpS/ASpS}}_{N\approx}}(\{\sigma+a\};\sigma)=P(a|\sigma)\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma)$$

$$\widehat{P^{\mathrm{SpS/ASpS}}_{N\approx}}(\{\sigma\};\sigma)=P(\tau|\sigma)\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma)$$

$$\forall s\in S_{\lessapprox\sigma}\setminus\{\sigma\},\widehat{P^{\mathrm{SpS/ASpS}}_{N\approx}}(\{s\};\sigma)=\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma)$$

The final sampling distributions for SpS and ASpS are given composing the neighbor distribution and the reference distribution:

$$P^{\mathrm{SpS/ASpS}}_{\mathrm{s}}(\cdot)=\sum_{\sigma\in S}P^{\mathrm{SpS/ASpS}}_{N\approx}(\cdot;\sigma)\widehat{P_{\mathrm{ref}}}(\sigma).$$

Note that SpS can be reformulated as follows:

$$\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma+b\};\sigma+a)=\begin{cases}\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma)\min(1,\frac{P(a|\sigma)}{P_{\mathrm{ref}}(a|\sigma)}) & \text{if } a=b,\\ \widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma)\frac{(1-\frac{P(a|\sigma)}{P_{\mathrm{ref}}(a|\sigma)})_+(P(b|\sigma)-P_{\mathrm{ref}}(b|\sigma))_+}{\sum_{z\in\Sigma(\sigma)}(P_{\mathrm{ref}}(z|\sigma)-P(z|\sigma))_+} & \text{if } a\neq b,\end{cases}$$

which resembles the solution for simple maximum coupling in Equation (1). Therefore, SpS applies maximum coupling step by step.

For both reference and target distributions, $m_{\mathrm{ref}}(a;\sigma)$ and $m(a;\sigma)$ denote the probability mass that can be used for coupling. Since $m^{\mathrm{ASpS}}_{\mathrm{ref}}(a;\sigma)\geq m^{\mathrm{SpS}}_{\mathrm{ref}}(a;\sigma)$, ASpS achieves a stronger coupling. As a result, both Theorems B.8 and B.9 illustrates ASpS's superior strength over SpS.

Lemma B.5. *Given a pseudo leaf-only reference distribution P_{ref} and one path $\sigma\in\mathrm{supp}(P_{\mathrm{ref}})$ such that $\Sigma(\sigma)\neq\phi$, for any $s\in S_{\lessapprox\sigma}$, the following holds:*

$$\sum_{a\in\Sigma(\sigma)}P_{\mathrm{ref}}(a|\sigma)P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(s;\sigma+a)=P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(s;\sigma).$$

Proof of Lemma B.5. We prove this lemma by considering two cases for s.

When $s=\sigma$, we have

$$\begin{aligned}&\sum_{a\in\Sigma(\sigma)}P_{\mathrm{ref}}(a|\,\sigma)P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma+a)\\ =&\sum_{a\in\Sigma(\sigma)}P_{\mathrm{ref}}(a|\,\sigma)\sum_{s\geq\sigma}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma+a)\\ =&\sum_{a\in\Sigma(\sigma)}P_{\mathrm{ref}}(a|\,\sigma)(\sum_{b\geq\Sigma(\sigma)}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma+b\};\sigma+a)+\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma+a))\\ =&\sum_{a\in\Sigma(\sigma)}\min(m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(a;\sigma),m^{\mathrm{SpS/ASpS}}(a;\sigma))\\ &+\sum_{a\in\Sigma(\sigma)}\sum_{b\geq\Sigma(\sigma),b\neq a}r^{\mathrm{SpS/ASpS}}(a;\sigma)(m^{\mathrm{SpS/ASpS}}(b;\sigma)-m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(b;\sigma))_+\end{aligned}$$

$$
\begin{aligned}
&= \sum_{a\in\Sigma(\sigma)} \min(m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(a;\sigma), m^{\mathrm{SpS/ASpS}}(a;\sigma)) \\
&\quad + \sum_{a\in\Sigma(\sigma)} \sum_{b\geq\Sigma(\sigma)} r^{\mathrm{SpS/ASpS}}(a;\sigma)(m^{\mathrm{SpS/ASpS}}(b;\sigma) - m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(b;\sigma))_+ \\
&\quad - \sum_{a\in\Sigma(\sigma)} r^{\mathrm{SpS/ASpS}}(a;\sigma)(m^{\mathrm{SpS/ASpS}}(a;\sigma) - m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(a;\sigma))_+ \\
&= \sum_{a\in\Sigma(\sigma)} \min(m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(a;\sigma), m^{\mathrm{SpS/ASpS}}(a;\sigma)) \\
&\quad + \sum_{b\geq\Sigma(\sigma)} (m^{\mathrm{SpS/ASpS}}(b;\sigma) - m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(b;\sigma))_+ \\
&\quad - \sum_{a\in\Sigma(\sigma)} \frac{(m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(a;\sigma) - m^{\mathrm{SpS/ASpS}}(a;\sigma))_+(m^{\mathrm{SpS/ASpS}}(a;\sigma) - m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(a;\sigma))_+}{\sum_{z\in\Sigma(\sigma)}(m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(z;\sigma) - m^{\mathrm{SpS/ASpS}}(z;\sigma))_+} \\
&= \sum_{a\in\Sigma(\sigma)} (\min(m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(a;\sigma), m^{\mathrm{SpS/ASpS}}(a;\sigma)) \\
&\qquad + (m^{\mathrm{SpS/ASpS}}(a;\sigma) - m_{\mathrm{ref}}^{\mathrm{SpS/ASpS}}(a;\sigma))_+) \\
&= \sum_{a\in\Sigma(\sigma)} m^{\mathrm{SpS/ASpS}}(a;\sigma) \\
&= \sum_{a\in\Sigma(\sigma)} P_{N\lessapprox}^{\widehat{\mathrm{SpS/AS}}\mathrm{pS}}(\{\sigma\};\sigma)P(a|\sigma) \\
&= P_{N\lessapprox}^{\widehat{\mathrm{SpS/AS}}\mathrm{pS}}(\{\sigma\};\sigma)
\end{aligned}
$$

For $s \in S_{\lessapprox\sigma} \setminus \{\sigma\}$, we have the following for SpS:

$$
\begin{aligned}
&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\,\sigma)\widehat{P_{N\lessapprox}^{\mathrm{SpS}}}(\{s\};\sigma+a) \\
=&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\,\sigma)\widehat{P_{N\lessapprox}^{\mathrm{SpS}}}(\{s\};\sigma) \\
=&P_{N\lessapprox}^{\widehat{\mathrm{SpS/ASpS}}}(\{s\};\sigma)
\end{aligned}
$$

For ASpS, we have:

$$
\begin{aligned}
&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\,\sigma)\widehat{P_{N\lessapprox}^{\mathrm{ASpS}}}(\{s\};\sigma+a) \\
=&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\,\sigma)\frac{r^{\mathrm{ASpS}}(a;\sigma)}{P_{\mathrm{ref}}(a|\sigma)}\widehat{P_{N\lessapprox}^{\mathrm{ASpS}}}(\{s\};\sigma) \\
=&\sum_{a\in\Sigma(\sigma)} r^{\mathrm{ASpS}}(a;\sigma)\widehat{P_{N\lessapprox}^{\mathrm{ASpS}}}(\{s\};\sigma) \\
=&\widehat{P_{N\lessapprox}^{\mathrm{ASpS}}}(\{s\};\sigma)
\end{aligned}
$$

□

Lemma B.6. *For a pseudo leaf-only reference distribution* P_{ref}*,* $\forall\sigma \in \mathrm{supp}(P_{\mathrm{ref}})$ *where* $\Sigma(\sigma) \neq \phi$*, and for all* $b \in \Sigma(\sigma)$*, we have:*

$$
\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)P_{N\lessapprox}^{\mathrm{SpS/ASpS}}(\sigma+b;\sigma+a) = P_{N\lessapprox}^{\mathrm{SpS/ASpS}}(\sigma;\sigma)P(b|\sigma).
$$

Proof of Lemma B.6.

$$
\begin{aligned}
&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\,\sigma)P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma+b;\sigma+a)\\
=&P_{\mathrm{ref}}(a|\,\sigma)P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma+a;\sigma+a)\\
&+\sum_{a\in\Sigma(\sigma),a\neq b} P_{\mathrm{ref}}(a|\,\sigma)P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma+b;\sigma+a)\\
=&\min(m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(a;\sigma),m^{\mathrm{SpS/ASpS}}(a;\sigma)\\
&+\sum_{a\in\Sigma(\sigma),a\neq b} r^{\mathrm{SpS/ASpS}}(a;\sigma)(m^{\mathrm{SpS/ASpS}}(b;\sigma)-m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(b;\sigma))_+\\
=&\min(m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(a;\sigma),m^{\mathrm{SpS/ASpS}}(a;\sigma)\\
&+\sum_{a\in\Sigma(\sigma)} r^{\mathrm{SpS/ASpS}}(a;\sigma)(m^{\mathrm{SpS/ASpS}}(b;\sigma)-m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(b;\sigma))_+\\
&-r^{\mathrm{SpS/ASpS}}(b;\sigma)(m^{\mathrm{SpS/ASpS}}(b;\sigma)-m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(b;\sigma))_+\\
=&\min(m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(a;\sigma),m^{\mathrm{SpS/ASpS}}(a;\sigma)\\
&+(m^{\mathrm{SpS/ASpS}}(b;\sigma)-m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(b;\sigma))_+\\
=&m^{\mathrm{SpS/ASpS}}(b;\sigma)\\
=&P^{\widehat{\mathrm{SpS/ASpS}}}_{N\lessapprox}(\{\sigma\};\sigma)P(b|\sigma)
\end{aligned}
$$

□

Lemma B.7. *Given a pseudo leaf-only reference distribution* P_{ref}*, for* $\forall\sigma\in\mathrm{supp}(P_{\mathrm{ref}})$*, and* $\forall s\in S_{\lessapprox\sigma}$*, we have:*

$$
\mathbb{E}_{\sigma'\sim\widehat{P_{\mathrm{ref}|\sigma}}}P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(s;\sigma')=P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(s;\sigma).
$$

Proof of Lemma B.7. Define the function $V(s')=P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(s;\sigma+s')$ for subsequences s'. This function satisfies the unbiased condition due to Lemma B.5. Applying the tree expectation theorem Theorem A.19, we can conclude that the expected value of the neighbor distribution calculated from the conditional distribution $P_{\mathrm{ref}|\sigma}$ is equal to the neighbor distribution at σ. □

Theorem B.8. *For every* $\sigma\in\mathrm{supp}(P_{\mathrm{ref}})$*, it holds that* $\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma)\geq\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma)$.

Proof of Theorem B.8. Initially, we have that $\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{\varepsilon\};\varepsilon)=\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\varepsilon\};\varepsilon)$.

Assume that $\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma)\geq\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma)$ holds. Given that

$$
P^{\widehat{\mathrm{SpS/ASpS}}}_{N\lessapprox}(\{\sigma+a\};\sigma+a)=\frac{\min(m^{\mathrm{SpS/ASpS}}_{\mathrm{ref}}(a;\sigma),m^{\mathrm{SpS/ASpS}}(a;\sigma))}{P_{\mathrm{ref}}(a|\sigma)},
$$

$$
m^{\mathrm{SpS}}_{\mathrm{ref}}(a;\sigma)=\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma)P_{\mathrm{ref}}(a|\sigma),
$$

$$
m^{\mathrm{ASpS}}_{\mathrm{ref}}(a;\sigma)=P_{\mathrm{ref}}(a|\sigma)\geq\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma)P_{\mathrm{ref}}(a|\sigma)=m^{\mathrm{SpS}}_{\mathrm{ref}}(a;\sigma).
$$

We have $\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{\sigma+a\};\sigma+a)\geq\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma+a\};\sigma+a)$. By applying mathematical induction, we completes the inductive step and the proof of the theorem. □

Theorem B.9. *If that the expected length of sequences sampled from the reference distribution is finite, i.e.,* $\mathbb{E}_{s\sim\widehat{P_{\mathrm{ref}}}}[\mathrm{len}(s)] \leq \infty$*, we have:*

$$\mathbb{E}_{s\sim\widehat{P^{\mathrm{ASpS}}}}[\mathrm{len}(s)] \geq \mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS}}}}[\mathrm{len}(s)].$$

Proof of Theorem B.9. Firstly, we have:

$$\mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS/ASpS}}_{N\approx}}(\cdot;\sigma)}[\mathrm{len}(s)] = \mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)] + \widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma)\mathbb{I}(\Sigma(\sigma)\neq\phi).$$

Due to Theorem B.8, we have$\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\{\sigma\};\sigma) \geq \widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\{\sigma\};\sigma)$. Therefore, we only need to prove the following:

$$\sum_{\sigma\in S}\mathbb{E}_{s\sim\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)]\widehat{P_{\mathrm{ref}}}(\sigma) \geq \sum_{\sigma\in S}\mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)]\widehat{P_{\mathrm{ref}}}(\sigma)$$

For σ such that $\Sigma(\sigma)\neq\phi$, we have:

$$\begin{aligned}
&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)\mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\cdot;\sigma+a)}[\mathrm{len}(s)] \\
=&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)\sum_{s\in S_{\approx\sigma}}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma+a)\,\mathrm{len}(s) \\
=&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)\sum_{s\in S_{\gtrsim\sigma}}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma+a)\,\mathrm{len}(s) \\
&+\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)\sum_{s\in S_{\lessapprox\sigma}}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma+a)\,\mathrm{len}(s) \\
=&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma+a)(\mathrm{len}(\sigma)+1) \\
&+\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)\sum_{s\in S_{\lessapprox\sigma}\setminus\{\sigma\}}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma+a)\,\mathrm{len}(s) \\
=&P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma)(\mathrm{len}(\sigma)+1) \\
&+\sum_{s\in S_{\lessapprox\sigma}\setminus\{\sigma\}}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma)\,\mathrm{len}(s) \\
=&P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma)+\sum_{s\in S_{\lessapprox\sigma}}\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\{s\};\sigma)\,\mathrm{len}(s) \\
=&P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma)+\mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS/ASpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)]
\end{aligned}$$

Define the difference in expected lengths when using ASpS and SpS as

$$V(\sigma) = \mathbb{E}_{s\sim\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)] - \mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)].$$

Then, for sequences σ where $\Sigma(\sigma)\neq\phi$, we see:

$$\begin{aligned}
&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)V(\sigma+a) \\
=&\sum_{a\in\Sigma(\sigma)} P_{\mathrm{ref}}(a|\sigma)(\mathbb{E}_{s\sim\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\cdot;\sigma+a)}[\mathrm{len}(s)] - \mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\cdot;\sigma+a)}[\mathrm{len}(s)])
\end{aligned}$$

$$
\begin{aligned}
&= P^{\mathrm{ASpS}}_{N\lessapprox}(\sigma;\sigma) - P^{\mathrm{SpS}}_{N\lessapprox}(\sigma;\sigma) + \mathbb{E}_{s\sim\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)] - \mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)] \\
&\geq V(\sigma)
\end{aligned}
$$

Utilizing the tree expectation theorem in Theorem A.19, we can deduce that:

$$
\sum_{\sigma\in S} V(\sigma)\widehat{P_{\mathrm{ref}}}(\sigma) \geq V(\varepsilon) = 0.
$$

This further implies that:

$$
\sum_{\sigma\in S} \mathbb{E}_{s\sim\widehat{P^{\mathrm{ASpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)]\widehat{P_{\mathrm{ref}}}(\sigma) \geq \sum_{\sigma\in S} \mathbb{E}_{s\sim\widehat{P^{\mathrm{SpS}}_{N\lessapprox}}(\cdot;\sigma)}[\mathrm{len}(s)]\widehat{P_{\mathrm{ref}}}(\sigma),
$$

which is precisely what we needed to prove. Therefore, the expected sequence length for ASpS is greater than or equal to that for SpS, which illustrates ASpS's superior strength. □

The following theorem guarantees that both SpS and ASpS can be applied recursively to converge to the target distribution correctly.

Theorem B.10. *Both SpS and ASpS constitute sub-samplers, i.e., $P_{\mathrm{s}}^{\mathrm{SpS/ASpS}} \preceq P$.*

Proof of Theorem B.10. As the target distribution P is a leaf-only distribution, $P(\tau|\sigma)$ is already minimized. We only need to prove:

$$
\forall\sigma \in \mathrm{supp}(P)\cap\mathrm{supp}(P_{\mathrm{ref}}),\ \forall a,b\in\Sigma(\sigma),\ P_{\mathrm{s}}^{\mathrm{SpS/ASpS}}(\sigma+a)P(\sigma+b) = P_{\mathrm{s}}^{\mathrm{SpS/ASpS}}(\sigma+b)P(\sigma+a).
$$

If $P_{\mathrm{ref}}(\tau|\sigma) = 1$, we have:

$$
\begin{aligned}
P_{\mathrm{s}}^{\mathrm{SpS/ASpS}}(\sigma+a) =& \sum_{\sigma'} P^{\mathrm{SpS/ASpS}}_{N\approx}(\sigma+a;\sigma')\widehat{P_{\mathrm{ref}}}(\{\sigma'\}) \\
=& P^{\mathrm{SpS/ASpS}}_{N\approx}(\sigma+a;\sigma)P_{\mathrm{ref}}(\sigma) \\
=& P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma)P(a|\sigma)P_{\mathrm{ref}}(\sigma),
\end{aligned}
$$

If $P_{\mathrm{ref}}(\tau|\sigma) = 0$, then:

$$
\begin{aligned}
P_{\mathrm{s}}^{\mathrm{SpS/ASpS}}(\sigma+a) =& \sum_{\sigma'} P^{\mathrm{SpS/ASpS}}_{N\approx}(\sigma+a;\sigma')\widehat{P_{\mathrm{ref}}}(\{\sigma'\}) \\
=& \sum_{\sigma''\in S_{\simeq\sigma+a}} \sum_{\sigma'\geq\sigma''} P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma+a;\sigma')\widehat{P_{\mathrm{ref}\,|\sigma''}}(\{\sigma'\})P_{\mathrm{ref}}(\sigma'') \\
=& \sum_{\sigma''\in S_{\simeq\sigma+a}} \mathbb{E}_{\sigma'\sim\widehat{P_{\mathrm{ref}\,|\sigma''}}}[P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma+a;\sigma')]P_{\mathrm{ref}}(\sigma'') \\
=& \sum_{\sigma''\in S_{\simeq\sigma+a}} P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma+a;\sigma'')P_{\mathrm{ref}}(\sigma'') \\
=& \sum_{b\in\Sigma(b)} P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma+a;\sigma+b)P_{\mathrm{ref}}(b|\sigma)P_{\mathrm{ref}}(\sigma) \\
=& P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma)P(a|\sigma)P_{\mathrm{ref}}(\sigma),
\end{aligned}
$$

In either cases, we have:

$$
\begin{aligned}
P_{\mathrm{s}}^{\mathrm{SpS/ASpS}}(\sigma+a)P(\sigma+b) =& P^{\mathrm{SpS/ASpS}}_{N\lessapprox}(\sigma;\sigma)P_{\mathrm{ref}}(\sigma)P(\sigma)P(a|\sigma)P(b|\sigma) \\
=& P_{\mathrm{s}}^{\mathrm{SpS/ASpS}}(\sigma+b)P(\sigma+a).
\end{aligned}
$$

Therefore, according to Theorem A.16, we establish that $P_{\mathrm{s}}^{\mathrm{SpS/ASpS}} \preceq P$. □

B.2. Application to LLM

For each LLM, there is a fixed token space Σ, within which there exists a special token known as the end-of-sequence (eos) token, $\mathrm{eos} \in \Sigma$. Use use a predicate, $\mathrm{terminated}(s)$ to represent whether the last token in the sequence s is the eos token. An LLM is a probability distribution as $\mathrm{LLM}(x_{n+1}|x_1, \ldots, x_n)$. A sequence can be continually extended by sampling and appending tokens until the eos token is reached, resulting in a terminated sequence.

A LLM constitutes a tree distribution. Note that all strings forms a tree space:

$$S = (\Sigma \setminus \{\mathrm{eos}\})^* \cup \{s + \mathrm{eos}\,|s \in (\Sigma \setminus \{\mathrm{eos}\})^*\},$$

$$\Sigma(s) = \begin{cases} \phi & \mathrm{terminated}(s), \\ \Sigma & \text{otherwise.} \end{cases}$$

The tree distribution is defined as:

$$P_{\mathrm{LLM}}(a|\sigma) = \begin{cases} 0 & \mathrm{terminated}(\sigma), \\ \mathrm{LLM}(a|\sigma) & \text{otherwise.} \end{cases}$$

$$P_{\mathrm{LLM}}(\tau|\sigma) = \begin{cases} 1 & \mathrm{terminated}(\sigma), \\ 0 & \text{otherwise.} \end{cases}$$

With this connection between TMC and LLM, we can apply ASpS to LLM as follows: 1. sample d reference tokens from the reference model that is less computationally expensive. 2. Calculate the neighbor distribution $P_{N\approx}^{\mathrm{ASpS}}(\cdot;\sigma)$. 3. Sample from the neighbor distribution to generate the final output.

C. Full Experiment Results

n		ATN			PTT		PPL
	SpS	ASpS	Improv. (%)	SpS	ASpS	Improv. (%)	Change (%)
1	1.284 ± 0.001	1.284 ± 0.001	0.0 ± 0.1	0.028 ± 0.000	0.028 ± 0.000	-0.1 ± 0.3	0.0 ± 2.5
2	1.381 ± 0.002	1.394 ± 0.002	0.9 ± 0.2	0.028 ± 0.000	0.028 ± 0.000	0.8 ± 0.4	-0.5 ± 2.4
3	1.458 ± 0.002	1.485 ± 0.002	1.8 ± 0.2	0.029 ± 0.000	0.028 ± 0.000	1.6 ± 0.4	0.9 ± 2.5
4	1.454 ± 0.002	1.489 ± 0.002	2.4 ± 0.2	0.031 ± 0.000	0.030 ± 0.000	2.1 ± 0.4	1.9 ± 2.5
5	1.436 ± 0.002	1.467 ± 0.002	2.1 ± 0.2	0.033 ± 0.000	0.033 ± 0.000	1.6 ± 0.4	0.7 ± 2.5
6	1.482 ± 0.002	1.507 ± 0.003	1.7 ± 0.2	0.034 ± 0.000	0.034 ± 0.000	1.0 ± 0.5	2.1 ± 2.4
7	1.477 ± 0.002	1.520 ± 0.003	2.9 ± 0.3	0.036 ± 0.000	0.036 ± 0.000	2.3 ± 0.5	0.8 ± 2.4
8	1.496 ± 0.003	1.525 ± 0.003	1.9 ± 0.3	0.038 ± 0.000	0.037 ± 0.000	1.2 ± 0.5	0.7 ± 2.5
9	1.482 ± 0.003	1.532 ± 0.003	3.3 ± 0.3	0.040 ± 0.000	0.039 ± 0.000	2.4 ± 0.5	-0.3 ± 2.5
10	1.468 ± 0.003	1.522 ± 0.003	3.7 ± 0.3	0.042 ± 0.000	0.041 ± 0.000	2.6 ± 0.5	1.0 ± 2.5

Table 6. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-7b model as target model and LLaMa-68m model as reference model on machine translation task.

n		ATN			PTT		PPL
	SpS	ASpS	Improv. (%)	SpS	ASpS	Improv. (%)	Change (%)
1	1.548 ± 0.001	1.548 ± 0.001	0.0 ± 0.1	0.023 ± 0.000	0.023 ± 0.000	-0.0 ± 0.1	0.0 ± 2.3
2	1.851 ± 0.001	1.869 ± 0.001	1.0 ± 0.1	0.021 ± 0.000	0.021 ± 0.000	0.8 ± 0.2	0.2 ± 2.3
3	2.006 ± 0.001	2.045 ± 0.002	1.9 ± 0.1	0.021 ± 0.000	0.020 ± 0.000	1.7 ± 0.2	-1.4 ± 2.3
4	2.125 ± 0.002	2.186 ± 0.002	2.9 ± 0.1	0.021 ± 0.000	0.020 ± 0.000	2.5 ± 0.2	3.8 ± 2.3
5	2.181 ± 0.002	2.270 ± 0.002	4.1 ± 0.1	0.022 ± 0.000	0.021 ± 0.000	3.7 ± 0.2	4.7 ± 2.3
6	2.204 ± 0.002	2.313 ± 0.002	5.0 ± 0.1	0.023 ± 0.000	0.022 ± 0.000	4.5 ± 0.2	-1.3 ± 2.3
7	2.223 ± 0.002	2.349 ± 0.003	5.7 ± 0.2	0.024 ± 0.000	0.023 ± 0.000	5.1 ± 0.3	0.3 ± 2.2
8	2.264 ± 0.002	2.386 ± 0.003	5.4 ± 0.2	0.025 ± 0.000	0.024 ± 0.000	4.8 ± 0.3	2.2 ± 2.3
9	2.238 ± 0.002	2.372 ± 0.003	6.0 ± 0.2	0.026 ± 0.000	0.025 ± 0.000	5.2 ± 0.3	0.4 ± 2.3
10	2.254 ± 0.002	2.412 ± 0.003	7.0 ± 0.2	0.027 ± 0.000	0.026 ± 0.000	6.1 ± 0.3	2.0 ± 2.3

Table 7. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-7b model as target model and LLaMa-68m model as reference model on text summarization task.

n		ATN			PTT		PPL
	SpS	ASpS	Improv. (%)	SpS	ASpS	Improv. (%)	Change (%)
1	1.487 ± 0.002	1.487 ± 0.002	0.0 ± 0.2	0.033 ± 0.000	0.033 ± 0.000	0.0 ± 0.2	0.0 ± 2.8
2	1.731 ± 0.003	1.748 ± 0.003	1.0 ± 0.3	0.030 ± 0.000	0.029 ± 0.000	0.9 ± 0.3	-1.6 ± 2.8
3	1.856 ± 0.004	1.904 ± 0.004	2.6 ± 0.3	0.029 ± 0.000	0.028 ± 0.000	2.4 ± 0.4	2.7 ± 2.9
4	1.920 ± 0.005	1.988 ± 0.005	3.5 ± 0.4	0.030 ± 0.000	0.029 ± 0.000	3.3 ± 0.4	-1.1 ± 2.9
5	1.942 ± 0.005	2.041 ± 0.006	5.1 ± 0.4	0.031 ± 0.000	0.029 ± 0.000	4.8 ± 0.5	2.9 ± 2.9
6	1.959 ± 0.005	2.074 ± 0.006	5.9 ± 0.4	0.032 ± 0.000	0.030 ± 0.000	5.4 ± 0.5	1.1 ± 2.9
7	1.977 ± 0.006	2.093 ± 0.007	5.8 ± 0.4	0.033 ± 0.000	0.031 ± 0.000	5.4 ± 0.5	3.8 ± 2.9
8	1.962 ± 0.006	2.107 ± 0.007	7.4 ± 0.5	0.034 ± 0.000	0.032 ± 0.000	6.8 ± 0.5	2.9 ± 3.0
9	1.976 ± 0.006	2.121 ± 0.007	7.3 ± 0.5	0.035 ± 0.000	0.033 ± 0.000	6.7 ± 0.5	0.3 ± 2.9
10	1.973 ± 0.006	2.112 ± 0.007	7.0 ± 0.5	0.037 ± 0.000	0.034 ± 0.000	6.4 ± 0.5	0.7 ± 2.8

Table 8. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-13b model as target model and LLaMa-68m model as reference model on open-ended text generation task.

n		ATN			PTT		PPL
	SpS	ASpS	Improv. (%)	SpS	ASpS	Improv. (%)	Change (%)
1	1.615 ± 0.002	1.615 ± 0.002	0.0 ± 0.2	0.035 ± 0.000	0.035 ± 0.000	0.0 ± 0.2	0.0 ± 3.0
2	1.987 ± 0.004	2.016 ± 0.004	1.5 ± 0.3	0.032 ± 0.000	0.031 ± 0.000	1.4 ± 0.3	-0.9 ± 3.2
3	2.233 ± 0.005	2.289 ± 0.005	2.5 ± 0.3	0.031 ± 0.000	0.030 ± 0.000	2.3 ± 0.4	-1.3 ± 3.0
4	2.373 ± 0.006	2.484 ± 0.007	4.7 ± 0.4	0.032 ± 0.000	0.031 ± 0.000	4.3 ± 0.5	0.7 ± 3.2
5	2.485 ± 0.007	2.621 ± 0.008	5.5 ± 0.5	0.033 ± 0.000	0.032 ± 0.000	5.2 ± 0.5	3.6 ± 3.0
6	2.534 ± 0.008	2.712 ± 0.010	7.0 ± 0.5	0.035 ± 0.000	0.033 ± 0.000	6.6 ± 0.5	-0.2 ± 2.9
7	2.536 ± 0.009	2.750 ± 0.010	8.5 ± 0.5	0.038 ± 0.000	0.035 ± 0.000	7.7 ± 0.6	-1.9 ± 2.9
8	2.568 ± 0.009	2.837 ± 0.011	10.5 ± 0.6	0.040 ± 0.000	0.036 ± 0.000	9.4 ± 0.6	-0.6 ± 2.9
9	2.585 ± 0.009	2.867 ± 0.012	10.9 ± 0.6	0.042 ± 0.000	0.038 ± 0.000	9.9 ± 0.6	-4.7 ± 2.9
10	2.591 ± 0.009	2.898 ± 0.012	11.8 ± 0.6	0.045 ± 0.000	0.040 ± 0.000	10.4 ± 0.6	0.8 ± 3.0

Table 9. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use Opt-13b model as target model and Opt-125m model as reference model on open-ended text generation task.

n		ATN			PTT		PPL
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.623 ± 0.002	1.623 ± 0.002	0.0 ± 0.2	0.022 ± 0.000	0.022 ± 0.000	−0.0 ± 0.2	0.0 ± 2.9
2	2.012 ± 0.004	2.036 ± 0.004	1.2 ± 0.3	0.021 ± 0.000	0.021 ± 0.000	1.1 ± 0.3	−0.6 ± 3.0
3	2.264 ± 0.005	2.328 ± 0.006	2.8 ± 0.3	0.022 ± 0.000	0.021 ± 0.000	2.7 ± 0.4	−1.9 ± 3.1
4	2.419 ± 0.007	2.539 ± 0.007	5.0 ± 0.4	0.023 ± 0.000	0.022 ± 0.000	4.7 ± 0.5	0.7 ± 3.0
5	2.520 ± 0.008	2.687 ± 0.009	6.6 ± 0.5	0.025 ± 0.000	0.023 ± 0.000	6.1 ± 0.5	−0.6 ± 3.0
6	2.581 ± 0.008	2.776 ± 0.010	7.5 ± 0.5	0.027 ± 0.000	0.025 ± 0.000	7.0 ± 0.5	−1.6 ± 2.9
7	2.598 ± 0.009	2.839 ± 0.011	9.3 ± 0.6	0.029 ± 0.000	0.027 ± 0.000	8.5 ± 0.6	−3.3 ± 3.0
8	2.644 ± 0.009	2.944 ± 0.012	11.3 ± 0.6	0.031 ± 0.000	0.028 ± 0.000	10.1 ± 0.6	−0.3 ± 3.0
9	2.673 ± 0.010	2.987 ± 0.013	11.7 ± 0.6	0.033 ± 0.000	0.030 ± 0.000	10.5 ± 0.6	−3.7 ± 3.0
10	2.677 ± 0.010	3.004 ± 0.013	12.2 ± 0.6	0.035 ± 0.000	0.032 ± 0.000	10.7 ± 0.6	−0.7 ± 3.0

Table 10. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use Opt-6.7b model as target model and Opt-125m model as reference model on open-ended text generation task.

n		ATN			PTT		PPL
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.615 ± 0.002	1.615 ± 0.002	0.0 ± 0.2	0.019 ± 0.000	0.019 ± 0.000	−0.1 ± 0.4	0.0 ± 3.5
2	2.004 ± 0.003	2.035 ± 0.004	1.5 ± 0.3	0.019 ± 0.000	0.019 ± 0.000	1.4 ± 0.4	−1.7 ± 3.5
3	2.254 ± 0.005	2.322 ± 0.005	3.0 ± 0.3	0.020 ± 0.000	0.019 ± 0.000	2.8 ± 0.5	−0.8 ± 3.4
4	2.420 ± 0.006	2.542 ± 0.007	5.0 ± 0.4	0.022 ± 0.000	0.021 ± 0.000	4.7 ± 0.5	0.0 ± 3.2
5	2.512 ± 0.007	2.692 ± 0.009	7.2 ± 0.5	0.024 ± 0.000	0.022 ± 0.000	6.6 ± 0.6	−0.7 ± 3.2
6	2.583 ± 0.008	2.830 ± 0.010	9.6 ± 0.5	0.026 ± 0.000	0.024 ± 0.000	8.7 ± 0.6	2.0 ± 3.2
7	2.613 ± 0.009	2.908 ± 0.011	11.3 ± 0.6	0.028 ± 0.000	0.025 ± 0.000	10.0 ± 0.6	−3.6 ± 3.4
8	2.647 ± 0.009	2.964 ± 0.012	12.0 ± 0.6	0.031 ± 0.000	0.027 ± 0.000	10.6 ± 0.6	−0.0 ± 3.3
9	2.598 ± 0.009	2.943 ± 0.012	13.3 ± 0.6	0.034 ± 0.000	0.030 ± 0.000	11.6 ± 0.7	2.8 ± 3.2
10	2.715 ± 0.010	3.098 ± 0.014	14.1 ± 0.7	0.035 ± 0.000	0.031 ± 0.000	12.2 ± 0.7	−4.2 ± 3.3

Table 11. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use GPT-Neo-2.7B model as target model and GPT-2 model as reference model on open-ended text generation task.

n		ATN			PTT		PPL
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.654 ± 0.002	1.654 ± 0.002	0.0 ± 0.1	0.023 ± 0.000	0.023 ± 0.000	0.0 ± 0.3	0.0 ± 3.3
2	2.075 ± 0.004	2.108 ± 0.004	1.6 ± 0.2	0.022 ± 0.000	0.022 ± 0.000	1.5 ± 0.4	−0.1 ± 3.3
3	2.373 ± 0.005	2.443 ± 0.006	2.9 ± 0.3	0.022 ± 0.000	0.022 ± 0.000	2.9 ± 0.5	−1.1 ± 3.3
4	2.559 ± 0.007	2.681 ± 0.007	4.8 ± 0.4	0.024 ± 0.000	0.023 ± 0.000	4.6 ± 0.5	0.4 ± 3.3
5	2.690 ± 0.008	2.881 ± 0.009	7.1 ± 0.5	0.025 ± 0.000	0.024 ± 0.000	6.6 ± 0.5	−1.0 ± 3.2
6	2.750 ± 0.009	3.037 ± 0.011	10.4 ± 0.5	0.027 ± 0.000	0.025 ± 0.000	9.3 ± 0.5	2.1 ± 3.2
7	2.811 ± 0.010	3.122 ± 0.012	11.1 ± 0.6	0.029 ± 0.000	0.026 ± 0.000	10.0 ± 0.6	−1.8 ± 3.3
8	2.839 ± 0.010	3.234 ± 0.013	13.9 ± 0.6	0.032 ± 0.000	0.028 ± 0.000	12.2 ± 0.6	−2.5 ± 3.4
9	2.876 ± 0.011	3.259 ± 0.014	13.3 ± 0.6	0.034 ± 0.000	0.030 ± 0.000	11.9 ± 0.6	1.2 ± 3.3
10	2.899 ± 0.011	3.313 ± 0.015	14.3 ± 0.7	0.036 ± 0.000	0.031 ± 0.000	12.2 ± 0.7	0.9 ± 3.3

Table 12. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use GPT-2-xl model as target model and GPT-2 model as reference model on open-ended text generation task.

n		**ATN**			**PTT**		**PPL**
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.498 ± 0.002	1.498 ± 0.002	0.0 ± 0.2	0.021 ± 0.000	0.021 ± 0.000	0.0 ± 0.4	0.0 ± 3.1
2	1.752 ± 0.003	1.768 ± 0.003	0.9 ± 0.3	0.020 ± 0.000	0.020 ± 0.000	0.8 ± 0.5	0.3 ± 3.0
3	1.872 ± 0.004	1.919 ± 0.004	2.5 ± 0.3	0.020 ± 0.000	0.019 ± 0.000	2.3 ± 0.5	−1.2 ± 2.9
4	1.959 ± 0.005	2.026 ± 0.005	3.4 ± 0.4	0.020 ± 0.000	0.020 ± 0.000	3.1 ± 0.5	−2.7 ± 3.0
5	1.997 ± 0.005	2.075 ± 0.006	3.9 ± 0.4	0.021 ± 0.000	0.021 ± 0.000	3.6 ± 0.6	−0.2 ± 3.0
6	2.001 ± 0.006	2.101 ± 0.006	5.0 ± 0.4	0.023 ± 0.000	0.021 ± 0.000	4.6 ± 0.6	−0.0 ± 3.0
7	2.009 ± 0.006	2.149 ± 0.007	7.0 ± 0.5	0.024 ± 0.000	0.022 ± 0.000	6.3 ± 0.6	1.7 ± 3.0
8	2.008 ± 0.006	2.169 ± 0.007	8.0 ± 0.5	0.025 ± 0.000	0.023 ± 0.000	7.2 ± 0.6	2.0 ± 3.1
9	2.015 ± 0.006	2.151 ± 0.007	6.8 ± 0.5	0.026 ± 0.000	0.025 ± 0.000	6.2 ± 0.6	−1.4 ± 3.1
10	2.003 ± 0.006	2.187 ± 0.008	9.2 ± 0.5	0.028 ± 0.000	0.025 ± 0.000	8.1 ± 0.6	1.0 ± 2.9

Table 13. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-7b model as target model and LLaMa-68m model as reference model on open-ended text generation task.

n		**ATN**			**PTT**		**PPL**
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.401 ± 0.002	1.401 ± 0.002	0.0 ± 0.2	0.023 ± 0.000	0.023 ± 0.000	−0.0 ± 0.4	0.0 ± 2.4
2	1.573 ± 0.003	1.576 ± 0.003	0.2 ± 0.2	0.022 ± 0.000	0.022 ± 0.000	0.1 ± 0.4	0.2 ± 2.4
3	1.639 ± 0.003	1.659 ± 0.003	1.2 ± 0.3	0.023 ± 0.000	0.023 ± 0.000	1.0 ± 0.5	−0.3 ± 2.1
4	1.657 ± 0.004	1.685 ± 0.004	1.7 ± 0.3	0.024 ± 0.000	0.024 ± 0.000	1.5 ± 0.5	−0.5 ± 2.3
5	1.683 ± 0.004	1.714 ± 0.004	1.8 ± 0.3	0.025 ± 0.000	0.025 ± 0.000	1.6 ± 0.5	−0.8 ± 2.4
6	1.693 ± 0.004	1.728 ± 0.004	2.1 ± 0.4	0.027 ± 0.000	0.026 ± 0.000	1.8 ± 0.5	−1.1 ± 2.3
7	1.698 ± 0.004	1.741 ± 0.005	2.5 ± 0.4	0.028 ± 0.000	0.027 ± 0.000	2.3 ± 0.5	−0.9 ± 2.3
8	1.690 ± 0.004	1.729 ± 0.004	2.3 ± 0.4	0.030 ± 0.000	0.029 ± 0.000	2.0 ± 0.5	−0.4 ± 2.4
9	1.694 ± 0.004	1.736 ± 0.005	2.5 ± 0.4	0.031 ± 0.000	0.030 ± 0.000	2.3 ± 0.5	1.0 ± 2.6
10	1.696 ± 0.004	1.737 ± 0.005	2.4 ± 0.4	0.033 ± 0.000	0.032 ± 0.000	2.1 ± 0.5	−1.7 ± 2.3

Table 14. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-2-7b model as target model and LLaMa-68m model as reference model on open-ended text generation task.

n		**ATN**			**PTT**		**PPL**
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.132 ± 0.001	1.132 ± 0.001	0.0 ± 0.1	0.029 ± 0.000	0.029 ± 0.000	−0.1 ± 0.2	0.0 ± 1.6
2	1.161 ± 0.001	1.164 ± 0.001	0.3 ± 0.1	0.031 ± 0.000	0.031 ± 0.000	0.2 ± 0.2	−0.5 ± 1.6
3	1.179 ± 0.001	1.182 ± 0.001	0.3 ± 0.1	0.033 ± 0.000	0.033 ± 0.000	0.1 ± 0.2	−0.6 ± 1.6
4	1.178 ± 0.001	1.181 ± 0.001	0.3 ± 0.1	0.035 ± 0.000	0.035 ± 0.000	0.1 ± 0.3	−0.9 ± 1.6
5	1.183 ± 0.001	1.187 ± 0.001	0.3 ± 0.1	0.037 ± 0.000	0.037 ± 0.000	0.0 ± 0.3	−0.5 ± 1.7
6	1.181 ± 0.001	1.187 ± 0.001	0.5 ± 0.1	0.040 ± 0.000	0.040 ± 0.000	0.2 ± 0.3	0.4 ± 1.7
7	1.181 ± 0.001	1.184 ± 0.001	0.2 ± 0.1	0.042 ± 0.000	0.042 ± 0.000	−0.1 ± 0.3	−0.6 ± 1.6
8	1.182 ± 0.001	1.188 ± 0.001	0.5 ± 0.2	0.044 ± 0.000	0.044 ± 0.000	0.2 ± 0.3	0.2 ± 1.7
9	1.184 ± 0.001	1.186 ± 0.001	0.2 ± 0.2	0.046 ± 0.000	0.046 ± 0.000	−0.1 ± 0.3	−0.6 ± 1.7
10	1.177 ± 0.001	1.185 ± 0.001	0.6 ± 0.2	0.048 ± 0.000	0.048 ± 0.000	0.3 ± 0.3	0.6 ± 1.7

Table 15. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-2-7b model as target model and LLaMa-68m model as reference model on machine translation task.

n		**ATN**			**PTT**		**PPL**
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.448 ± 0.001	1.448 ± 0.001	0.0 ± 0.1	0.024 ± 0.000	0.024 ± 0.000	−0.0 ± 0.2	0.0 ± 0.3
2	1.662 ± 0.001	1.666 ± 0.001	0.2 ± 0.1	0.023 ± 0.000	0.023 ± 0.000	0.2 ± 0.2	0.1 ± 0.3
3	1.761 ± 0.001	1.769 ± 0.001	0.5 ± 0.1	0.023 ± 0.000	0.023 ± 0.000	0.4 ± 0.2	0.0 ± 0.3
4	1.817 ± 0.001	1.827 ± 0.001	0.5 ± 0.1	0.024 ± 0.000	0.024 ± 0.000	0.4 ± 0.2	−0.2 ± 0.3
5	1.843 ± 0.002	1.857 ± 0.002	0.7 ± 0.1	0.025 ± 0.000	0.025 ± 0.000	0.5 ± 0.2	0.4 ± 0.3
6	1.851 ± 0.002	1.867 ± 0.002	0.8 ± 0.1	0.026 ± 0.000	0.026 ± 0.000	0.7 ± 0.2	0.0 ± 0.3
7	1.859 ± 0.002	1.870 ± 0.002	0.6 ± 0.1	0.028 ± 0.000	0.028 ± 0.000	0.4 ± 0.2	0.0 ± 0.3
8	1.865 ± 0.002	1.882 ± 0.002	0.9 ± 0.1	0.029 ± 0.000	0.029 ± 0.000	0.7 ± 0.2	−0.0 ± 0.3
9	1.867 ± 0.002	1.884 ± 0.002	0.9 ± 0.1	0.031 ± 0.000	0.030 ± 0.000	0.6 ± 0.2	0.1 ± 0.3
10	1.874 ± 0.002	1.889 ± 0.002	0.8 ± 0.1	0.032 ± 0.000	0.032 ± 0.000	0.5 ± 0.2	0.2 ± 0.3

Table 16. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-2-7b model as target model and LLaMa-68m model as reference model on text summarization task.

n		**ATN**			**PTT**		**PPL**
	SpS	**ASpS**	**Improv. (%)**	**SpS**	**ASpS**	**Improv. (%)**	**Change (%)**
1	1.384 ± 0.002	1.384 ± 0.002	0.0 ± 0.2	0.035 ± 0.000	0.035 ± 0.000	−0.1 ± 0.2	0.0 ± 2.1
2	1.539 ± 0.003	1.546 ± 0.003	0.5 ± 0.2	0.033 ± 0.000	0.033 ± 0.000	0.5 ± 0.3	0.2 ± 2.3
3	1.604 ± 0.003	1.617 ± 0.003	0.8 ± 0.3	0.034 ± 0.000	0.033 ± 0.000	0.7 ± 0.4	−1.5 ± 2.2
4	1.625 ± 0.003	1.652 ± 0.004	1.6 ± 0.3	0.035 ± 0.000	0.034 ± 0.000	1.4 ± 0.4	0.1 ± 2.3
5	1.647 ± 0.004	1.680 ± 0.004	2.0 ± 0.3	0.036 ± 0.000	0.035 ± 0.000	1.9 ± 0.4	0.0 ± 2.3
6	1.645 ± 0.004	1.679 ± 0.004	2.0 ± 0.3	0.038 ± 0.000	0.037 ± 0.000	1.8 ± 0.5	−0.6 ± 2.4
7	1.649 ± 0.004	1.693 ± 0.004	2.6 ± 0.3	0.039 ± 0.000	0.038 ± 0.000	2.4 ± 0.5	−0.7 ± 2.3
8	1.642 ± 0.004	1.677 ± 0.004	2.2 ± 0.3	0.041 ± 0.000	0.040 ± 0.000	1.9 ± 0.4	−1.6 ± 2.4
9	1.647 ± 0.004	1.678 ± 0.004	1.9 ± 0.3	0.042 ± 0.000	0.042 ± 0.000	1.7 ± 0.5	−0.9 ± 2.3
10	1.651 ± 0.004	1.682 ± 0.004	1.9 ± 0.4	0.044 ± 0.000	0.043 ± 0.000	1.6 ± 0.5	0.7 ± 2.3

Table 17. Summary of accepted token numbers per step (ATN) and per token time (PTN) improvement and perplexity (PPL) change for ASpS over SpS. n is the number of reference token. We use LLaMa-2-13b model as target model and LLaMa-68m model as reference model on open-ended text generation task.