

Autoregressive Decoding as a Martingale: Characterizing the Stochastic Dynamics of LLM Generation

Xiahui Zhu^{1*}, Zhenyi Qi^{1*}, Wanwei Liu¹, Haoang Chi^{1†}, Yang Tan¹, Xinyu Sun¹, Ji Wang¹

¹College of Computer, National University of Defense Technology, Changsha, Hunan, China
zhuxiahui21@nudt.edu.cn

Abstract

Large language models (LLMs) have achieved remarkable success across a wide range of language tasks. At their core, autoregressive LLMs are conditional probability models that learn to approximate the next-token distribution given the preceding context. Despite this probabilistic nature, the stochastic dynamics of autoregressive decoding have rarely been investigated from a rigorous probabilistic perspective. In this work, we define the per-step prediction error as the difference between the sampled token indicator and the model’s conditional predictive distribution, and prove that the cumulative error process is a martingale with respect to the decoding filtration, implying that its conditional first-order moment is identically zero. Beyond this first-order property, we derive a closed-form expression for the second-order behavior of the martingale sequence: its predictable quadratic variation reduces to $\sum_t (1 - p(y_t | y_{<t}))$, computable from realized token probabilities alone. This identity establishes a direct theoretical connection between sequence fluctuations and predictive uncertainty during autoregressive decoding. White-box experiments on Llama-3.x and Qwen2.5 families (1B to 72B parameters) verify that the empirically measured second-order moments closely match the theoretical predictions derived from our martingale analysis. These results provide a probabilistic framework for analyzing autoregressive LLM decoding, reveal intrinsic probabilistic fingerprints of different language models, and establish a rigorous mathematical foundation for LLM-generated text detection and uncertainty quantification.

Introduction

Large Language Models (LLMs) have achieved remarkable success across a wide range of natural language tasks through autoregressive decoding (Touvron et al. 2023; Achiam et al. 2023), where tokens are sequentially sampled from context-dependent predictive distributions. Despite their probabilistic nature, the stochastic dynamics governing the evolution of LLM generation sequences remain insufficiently understood from a rigorous probabilistic perspective (Welleck et al. 2020).

Since autoregressive LLMs are fundamentally conditional probability models, understanding their generation dynamics is a natural yet underexplored direction. Existing analyses mainly focus on local predictive behaviors, including likelihood, uncertainty, entropy, and calibration metrics (Holtzman et al. 2020; Meister, Cotterell, and Vieira 2020; Guo

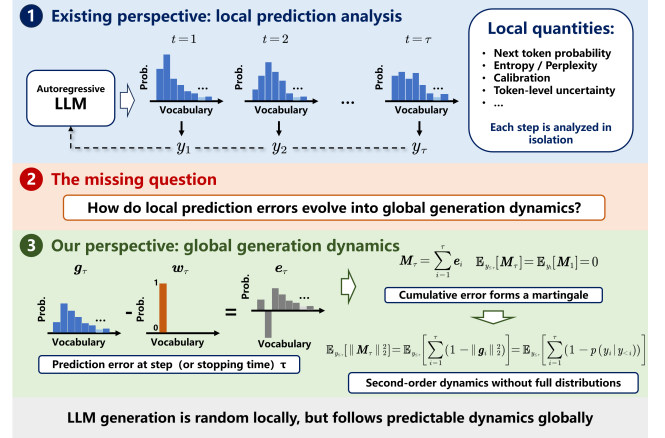


Figure 1: **A stochastic process perspective of autoregressive generation.** Existing analyses focus on local token-level uncertainty, while the evolution of prediction errors across the entire generation sequence remains unclear. Our perspective models autoregressive decoding as a stochastic process, where cumulative prediction errors form a martingale and exhibit predictable second-order dynamics without requiring full prediction distributions.

et al. 2017; Desai and Durrett 2020; Gal and Ghahramani 2016). However, these approaches characterize individual prediction distributions rather than the evolution of stochastic deviations over complete decoding sequences.

A fundamental question therefore remains open: how do stochastic prediction errors evolve during autoregressive generation? Since the predictive distribution at each step depends on previously generated tokens, a principled stochastic-process characterization is needed to understand the intrinsic randomness of language generation.

In this work, we introduce a martingale perspective to analyze autoregressive LLM decoding. We define token-level prediction error as the difference between sampled token indicators and conditional predictive probabilities, and prove that the cumulative error process forms a martingale with respect to the decoding filtration. This establishes a fundamental first-order zero-drift property: although individual token realizations are stochastic, their cumulative deviation

*These authors contributed equally.

†Corresponding author.

remains zero in expectation.

Beyond the first-order property, we characterize the fluctuation dynamics through second-order analysis. We derive an analytical identity showing that the martingale second-order moment can be computed directly from realized token probabilities, avoiding explicit construction of high-dimensional error vectors. This establishes a theoretical connection between sequence-level stochastic fluctuations and predictive distributions in autoregressive decoding.

To validate the theoretical framework, we conduct white-box experiments on multiple LLM families using Monte Carlo sampling. Results show that empirical decoding trajectories closely follow the theoretically predicted first- and second-order behaviors.

The main contributions of this work are summarized as follows:

- We formulate autoregressive LLM decoding as a stochastic process and establish that the cumulative prediction error vector forms a martingale with respect to the decoding filtration.
- We derive first- and second-order moment characterizations of the decoding martingale, including an analytical quadratic variation expression determined by the model’s predictive probability distribution.
- We provide comprehensive white-box empirical verification across multiple modern LLM families and scales, confirming the generality of the derived stochastic properties.

Related Work

Analysis of Large Language Model Generation Dynamics. Understanding the behavior of autoregressive generation has become an important research direction for large language models. Previous studies have mainly investigated generation quality, uncertainty estimation, degeneration phenomena, and decoding strategies (Fan, Lewis, and Dauphin 2018; Li et al. 2023; Leviathan, Kalman, and Matias 2023; Mitchell et al. 2023). Holtzman et al. analyzed neural text degeneration and proposed nucleus sampling to alleviate unreliable low-probability token selection during generation (Holtzman et al. 2020; Meister et al. 2022). Meister et al. studied the relationship between beam search and probabilistic sequence generation, highlighting the importance of decoding strategies in controlling model behavior (Meister, Cotterell, and Vieira 2020).

Beyond decoding algorithms, recent studies have explored uncertainty estimation and confidence calibration for neural language models. Token-level likelihood, entropy, calibration metrics, and semantic-level uncertainty estimation approaches have been widely adopted to quantify predictive confidence during generation (Guo et al. 2017; Jiang et al. 2021; Kuhn, Gal, and Farquhar 2023; Lin, Hilton, and Evans 2022; Kadavath et al. 2022; Brier 1950). However, these approaches primarily focus on the marginal properties of individual predictions. They characterize uncertainty at isolated decoding steps but do not explicitly model how stochastic deviations accumulate and evolve along the autoregressive

generation sequence. In contrast, our work studies the sequential stochastic dynamics of prediction errors and establishes a martingale-based characterization of autoregressive decoding.

Theoretical Understanding of Neural Language Models.

Recent advances in large language models have motivated increasing efforts toward developing theoretical explanations of their capabilities and behaviors (Wei et al. 2022; Ouyang et al. 2022). Existing theoretical studies have investigated scaling behavior, optimization properties, and internal mechanisms of large neural language models (Garg et al. 2022). Kaplan et al. revealed empirical scaling laws describing the relationship between model size, dataset size, and performance (Kaplan et al. 2020). Hoffmann et al. further refined these observations by studying compute-optimal scaling strategies for large language models (Hoffmann et al. 2022).

In the realm of interpretability, current mechanistic approaches have made significant strides in uncovering how transformer architectures internally represent and process information (Meng et al. 2022; Wang et al. 2023; Olsson et al. 2022). While these studies yield profound insights into model parameters and internal activations, the stochastic evolution of the generated sequence over time remains largely under-explored. Our work provides a complementary perspective by characterizing the macroscopic probabilistic dynamics induced during autoregressive inference, entirely independent of specific architectural components.

Martingale Theory in Machine Learning. Martingale theory provides a fundamental mathematical framework for analyzing stochastic processes whose future conditional expectation is determined by their current information state. Classical martingale results have played an important role in probability theory, sequential analysis, and stochastic optimization (Doob 1953; Freedman 1975; McDiarmid 1989). In machine learning, martingale techniques have been widely applied to derive concentration bounds, analyze online learning algorithms, and establish theoretical guarantees for sequential decision-making problems. For example, Azuma-Hoeffding inequalities and related martingale concentration tools have been extensively used in reinforcement learning and multi-armed bandit analysis (Azuma 1967; Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2020).

Stochastic Process Perspectives in Machine Learning.

Stochastic process theory has provided powerful tools for analyzing sequential behaviors in machine learning systems. Recent studies have employed stochastic differential equations, diffusion processes, and stochastic optimization frameworks to characterize the dynamics of learning algorithms and generative models (Song et al. 2021; Li et al. 2022). However, these analyses mainly focus on training dynamics or continuous generative processes, while the discrete stochastic evolution of autoregressive token generation remains largely unexplored. Our work differs by directly modeling the inference-time decoding sequence of LLMs as a discrete martingale process.

Despite its importance in sequential learning theory, mar-

tingale analysis has not been systematically explored for autoregressive language generation. In this work, we bridge this gap by introducing a martingale formulation for the cumulative prediction error process of LLM decoding. This perspective enables rigorous first- and second-order moment analysis of the generation sequence and provides a probabilistic characterization of autoregressive decoding dynamics.

Preliminaries

In this section, we briefly review the central concepts of martingale theory within the context of probability theory and stochastic processes, providing a theoretical underpinning for the modeling of the autoregressive generation process discussed in the subsequent chapter.

Consider a complete probability space $(\Omega, \mathcal{F}, P)$. In the context of discrete-time autoregressive text generation, the output token sequence $\{y_t\}_{t \geq 1}$ produced by the stochastic process corresponds to the observed sequences in the sample space.

Definition 1 (Filtration). A sequence of sub- σ -algebras $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 1}$ on the measure space is called a *filtration* if it satisfies the property of monotonicity, i.e., $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}_n \subseteq \mathcal{F}$. In this study, we define $\mathcal{F}_t = \sigma(y_1, y_2, \dots, y_t)$ as the σ -algebra generated by the historical token sequence up to time t , representing all observed historical text information (Doob 1953).

Definition 2 (Martingale). Given a filtration $\mathbb{F}$, a stochastic process $M = \{M_t\}_{t \geq 1}$ adapted to this filtration is called a martingale with respect to $\mathbb{F}$ if it satisfies:

1. **Integrability:** For any $t \geq 1$, $\mathbb{E}[|M_t|] < \infty$;
2. **Martingale Property:** For any $t \geq 1$, the conditional expectation equality holds almost surely:

$$\mathbb{E}[M_{t+1} | \mathcal{F}_t] = M_t. \quad (1)$$

In the context of text generation, this property implies that if the model prediction is accurate and unbiased, then the expected increment of the future evolution of the cumulative prediction error—where observed values deviate from predicted probabilities—given the historical information, should be zero.

In actual text generation tasks, the length of a sequence is often not fixed but determined by specific stopping conditions (such as generating an [EOS] token or reaching a maximum length limit). This is mathematically characterized by “stopping time”.

Definition 3 (Stopping Time). A random variable $\tau : \Omega \rightarrow \mathbb{N} \cup \{+\infty\}$ is called a stopping time with respect to filtration $\mathbb{F}$ if for any t , the event $\{\tau = t\}$ depends only on the historical information up to time t , i.e., $\{\tau = t\} \in \mathcal{F}_t$ (Doob 1953).

With the definition of stopping time, we have the theorem as follows.

Theorem 1 (Martingale Stopping Theorem). *Let $M = \{M_t\}_{t \geq 1}$ be a martingale with respect to $\mathbb{F}$, and τ be a stopping time. If τ is bounded (i.e., there exists a constant K such that $\tau \leq K$ almost surely), then the expected value of the process at the stopping time equals its expected value at the initial time: $\mathbb{E}[M_\tau] = \mathbb{E}[M_1]$.*

Noticing that for LLMs the maximum generation length ensures a bounded stopping time, MST conditions are always satisfied. This guarantees that the expected cumulative prediction error remains statistically conserved for model-generated sequences, independent of stochastic termination.

Theoretical Analysis

This section aims to prove and explore martingale properties of LLMs. First, we formalize the generation process of autoregressive language models as a stochastic process, prove that it satisfies the martingale property, and further examine the properties of the first-order moment. Second, based on previous results, we turn our attention to the second-order moment of the martingale, where a series of formulas are given to accommodate the practical operation of LLMs. Finally, we examine the meaning of martingale second-order moments, linking it to the stability of model outputs.

Notations

Let $\mathbf{X}$ and $\mathbf{Y} = \{y_t\}_{1 \leq t \leq l_{\max}}$ be a discrete sequence of input and output text respectively, where $y_t \in \mathcal{V}$ denotes the token at time step t , $l_{\max}$ denotes the maximum generation length of LLM, and $\mathcal{V}$ denotes the complete vocabulary set of the model. We define the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 1}$, where $\mathcal{F}_t = \sigma(y_1, y_2, \dots, y_t)$. To be precise, $\mathcal{F}_t$ is the power set of $\mathcal{V}^t$.

The generation process of LLM, an autoregressive model, is essentially a parameterized conditional probability distribution $p(y_t | y_{<t}; \mathbf{X}, \theta)$, often abbreviated as $p(y_t | y_{<t})$ when $\mathbf{X}$ and θ remain unchanged and are not the focus of discussion, where $y_{<t}$ represents the sequence $y_1 y_2 \dots y_{t-1}$. Similarly, we denote $y_{\leq t}$ as the sequence $y_1 y_2 \dots y_t$.

In what follows, the notation $\mathbb{E}_{y_t}[\cdot]$, $\mathbb{E}_{y_{<t}}[\cdot]$ will denote the expectations taken over the random variables y_t and $y_1, y_2, \dots, y_{t-1}$ respectively. Other cases follow analogously.

In practice, large language models output vectors rather than scalars, so we introduce the following vectors in $\mathbb{R}^{|\mathcal{V}|}$:

Predictive Probability Vector (g_t):

$$g_{t,y} = p(y | y_{<t}; \mathbf{X}, \theta). \quad (2)$$

This vector is just the predictive probability vector given by the model for next token. It depends on input $\mathbf{X}$, parameters θ and the history of output $y_{<t}$.

Token Indicator Vector (w_t):

$$w_{t,y} = \mathbb{I}_{\{y=y_t\}}. \quad (3)$$

It's the one-hot vector for the token actually generated by the model at time t .

Prediction Error Vector (e_t):

$$e_t = w_t - g_t. \quad (4)$$

Notice that $\mathbb{E}_{y_t}[w_t | y_{<t}] = g_t$, error vector e_t actually measures the discrepancy between w_t and its expectation g_t .

Cumulative Prediction Error Vector (M_t):

$$M_t = \sum_{i=1}^t e_i. \quad (5)$$

Just as the name suggests, M_t is the accumulation of error e_t , which is the central focus of our paper. We will proof that it is actually a martingale and elaborate on its various properties.

To more meticulously model the stochastic process underlying text generation by LLM, it is essential to account for the randomness in the length of the generated text. Hence, we introduce the concept of stopping time:

Stopping Time (τ): A random variable $\tau : \Omega \rightarrow \{1, 2, \dots, l_{\max}\}$ which is defined as follows:

$$\tau := \inf \{t > 0 : y_t \text{ is [EOS]}\} \cup \{l_{\max}\}. \quad (6)$$

In the context of token tracing, τ is actually the output length. When [EOS] is generated, τ is the first occurrence time of the end-of-sequence token. Otherwise it is the maximum generation length $l_{\max}$ with length limit reached. It can be easily verified that $\{\tau = t\} \in \mathcal{F}_t$ and thus τ is a stopping time with respect to $\mathbb{F}$.

It should be noted that τ is strictly bounded by the maximum generation length $l_{\max}$, which satisfies the boundedness condition required for martingale stopping theorem.

Martingale and its Property of First-Order Moment

The cornerstone of our entire theoretical analysis rests upon the fact that M_t is a martingale with respect to the filtration $\mathbb{F}$.

Theorem 2. *The stochastic process $\{M_t\}_{t \geq 1}$ is a martingale with respect to the filtration $\mathbb{F}$.*

Furthermore, we define the first-order moment of the martingale as one aspect of the study of its martingale properties.

Definition 4 (First-Order Moment of the Martingale). Given a martingale $\{M_t\}_{t \geq 1}$, the first-order moment of martingale $\{M_t\}_{t \geq 1}$ at time t (or stopping time τ) is $\mathbb{E}_{y_{\leq t}}[M_t]$ (or $\mathbb{E}_{y_{\leq l_{\max}}}[M_\tau]$).

It is not difficult to see that the first-order moment of the martingale M_t in fact characterizes the expected value of the accumulated prediction error. By the martingale expectation property, it follows that

$$\mathbb{E}_{y_{\leq t}}[M_t] = \mathbb{E}_{y_1}[M_1] = \mathbb{E}_{y_1}[w_1] - g_1 = 0. \quad (7)$$

Furthermore, by martingale stopping theorem, we obtain

$$\mathbb{E}_{y_{\leq l_{\max}}}[M_\tau] = \mathbb{E}_{y_1}[M_1] = 0. \quad (8)$$

It should be emphasized that equation (8) holds for arbitrary $\mathbb{F}$ -stopping times, thereby rendering it a stronger version of equation (7)—indeed, equation (7) is obtained as the special case where $\tau \equiv t$. In what follows, any theorem or equation stated in terms of stopping times is uniformly understood as a strengthened formulation of the corresponding non-stopping-time version; we shall not repeat this remark subsequently.

Equations (7) and (8) show that whether for fixed time t or stopping time τ , the first-order moment of $\{M_t\}_{t \geq 1}$ is always 0, which means accumulated prediction error vanishes to 0 in the long run. This implies that the first-order moment alone cannot furnish sufficient information for analytical and practical purposes, which motivates us to investigate the properties of the second-order moment in the next section.

Calculation of the Second-Order Moment of the Martingale

To gain a deeper characterization of martingale properties, we now introduce the second-order moment of the martingale.

Definition 5 (Second-Order Moment of the Martingale). Given a martingale $\{M_t\}_{t \geq 1}$, the second-order moment of martingale $\{M_t\}_{t \geq 1}$ at time t (or stopping time τ) is $\mathbb{E}[\|M_t\|_2^2]$ (or $\mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2]$).

Regarding the calculation of the second-order moment, we present the following theorem:

Theorem 3. *The second-order moment of martingale $\{M_t\}_{t \geq 1}$ can be expressed as follows:*

$$\mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2] = \mathbb{E}_{y_{\leq t}} \left[\sum_{i=1}^t (1 - \|g_i\|_2^2) \right]. \quad (9)$$

Theorem 3 offers a viable approach for computing the second order moment, which is determined exclusively by the vector g_t , indicating that it constitutes an inherent characteristic of the LLM. However, it doesn't consider the random generation text length τ —it can only calculate for fixed length t . Hence, we turn to the formula for the second order moment associated with stopping time.

Theorem 4. *The second order moment of martingale $\{M_t\}_{t \geq 1}$ with stopping time can be expressed as follows:*

$$\mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2] = \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} (1 - \|g_t\|_2^2) \right]. \quad (10)$$

In contrast to Theorem 3, Theorem 4 better reflects the practical operation of LLMs. Nevertheless, it presents an application-level concern: the individual components of g_t are not necessarily all available, depending on the accessibility (open-source vs. closed-source) of the model. In response, we present the following theorem:

Theorem 5. *The second-order moment of martingale $\{M_t\}_{t \geq 1}$ can be expressed as follows:*

$$\mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2] = \mathbb{E}_{y_{\leq t}} \left[\sum_{i=1}^t (1 - p(y_i | y_{<i})) \right]. \quad (11)$$

Theorem 6. *The second order moment of martingale $\{M_t\}_{t \geq 1}$ with stopping time can be expressed as follows:*

$$\begin{aligned} & \mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2] \\ &= \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} (1 - p(y_t | y_{<t})) \right]. \end{aligned} \quad (12)$$

Compared with equations (9) and (10), equations (11) and (12) do not require knowledge of every component of g_t . They only need the actually generated tokens $y_{\leq t}$ and their associated conditional probabilities $p(y_t | y_{<t})$, which are usually accessible from typical LLM inference APIs. In addition, their computation does not require squaring and summing over $|\mathcal{V}|$ components, but only scalar summations. This suggests that equations (11) and (12) are both more computationally efficient and easier to implement in practice.

Second-Order Moment and Stability of Model Outputs

In fact, the second order moment of martingales is closely tied to the stability of LLM outputs. Here we restrict our attention to syntactic stability, which refers to the similarity of generated texts across repeated generating.

Theorem 7. *The second-order moment of martingale (with stopping time) can be expressed as follows:*

$$\mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2] = \mathbb{E}_{y_{\leq t}}\left[\sum_{i=1}^t e_i^2\right], \quad (13)$$

$$\mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2] = \mathbb{E}_{y_{\leq l_{\max}}}\left[\sum_{t=1}^\tau e_t^2\right]. \quad (14)$$

Theorem 7 implies that the second order moment of martingales captures the expected sum of squared errors at each step, which essentially quantifies the uncertainty in LLM outputs. The larger this value, the less reproducible the model is, and vice versa. It equals zero if and only if the error at every step is zero—that is, the model’s generation is entirely determined by the given input and parameters.

The following theorem further characterizes the relationship between the second order moment of martingales and the output stability of the model.

Theorem 8. *For the second order moments $\mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2]$ and $\mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2]$, the following relation holds:*

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) = \mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2] + \sum_{i=1}^t \text{Var}(\mathbf{g}_i), \quad (15)$$

$$\begin{aligned} & \mathbb{E}_{y_{\leq l_{\max}}}\left[\sum_{t=1}^\tau \text{Var}(\mathbf{w}_t)\right] \\ &= \mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2] + \mathbb{E}_{y_{\leq l_{\max}}}\left[\sum_{t=1}^\tau \text{Var}(\mathbf{g}_t)\right]. \end{aligned} \quad (16)$$

Then

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) \geq \mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2], \quad (17)$$

$$\mathbb{E}_{y_{\leq l_{\max}}}\left[\sum_{t=1}^\tau \text{Var}(\mathbf{w}_t)\right] \geq \mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2]. \quad (18)$$

Theorem 8 states that the sum of variances of $\mathbf{w}_i$ admits a decomposition into two components: the second order moment of the martingale and the sum of variances of $\mathbf{g}_i$, and the second order moment serves as its lower bound. Specifically, the second order moment measures the randomness within the same output sequence, while the sum of variances of $\mathbf{g}_i$ measures the randomness across different output sequences. It can be shown that minimizing the second order moment of the martingale is equivalent to minimizing the sum of variances of $\mathbf{w}_i$.

Theorem 9. *The following equivalences hold:*

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) = 0 \iff \mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2] = 0, \quad (19)$$

$$\begin{aligned} \mathbb{E}_{y_{\leq l_{\max}}}\left[\sum_{t=1}^\tau \text{Var}(\mathbf{w}_t)\right] = 0 &\iff \\ \mathbb{E}_{y_{\leq l_{\max}}}[\|M_\tau\|_2^2] = 0. \end{aligned} \quad (20)$$

With Theorem 9, we naturally ask whether the maximization of the second order moment of the martingale is equivalent to the maximization of the sum of variances of $\mathbf{w}_i$. The answer is negative. We have the following theorem:

Theorem 10. *The following equivalences hold:*

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) = t(1 - \frac{1}{|\mathcal{V}|}) \iff \quad (21)$$

$$\forall 1 \leq i \leq t : \mathbb{E}_{y_{< i}}[\mathbf{g}_i] = \frac{1}{|\mathcal{V}|}\mathbf{1},$$

$$\mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2] = t(1 - \frac{1}{|\mathcal{V}|}) \iff \quad (22)$$

$$\forall 1 \leq i \leq t : (\mathbf{g}_i = \frac{1}{|\mathcal{V}|}\mathbf{1} \text{ a.e.}),$$

where $\mathbf{1}$ is the all-one vector.

The following implication hold:

$$\begin{aligned} \mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2] = t(1 - \frac{1}{|\mathcal{V}|}) &\Rightarrow \\ \sum_{i=1}^t \text{Var}(\mathbf{w}_i) = t(1 - \frac{1}{|\mathcal{V}|}). \end{aligned} \quad (23)$$

Counterexample to the converse of equation (23) is provided, as shown in Figure 2.

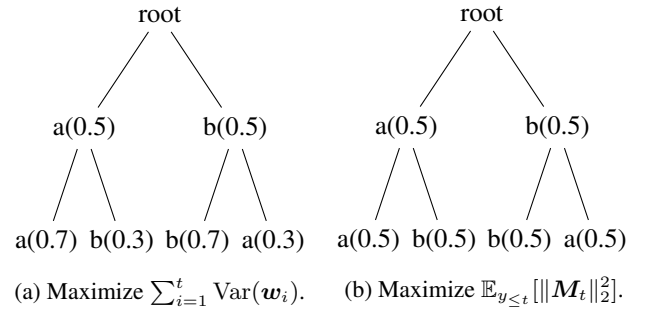


Figure 2: The case where $\mathbb{E}_{y_{\leq t}}[\|M_t\|_2^2]$ isn’t maximized when $\sum_{i=1}^t \text{Var}(\mathbf{w}_i)$ is maximized.

In Figures 2a and 2b, each node has $|\mathcal{V}|$ child nodes, representing the tokens that can be generated at the next time step, with each child node associated with a selection probability $p(y_t | y_{< t})$.

For Figure 2a, we have

$$\begin{aligned} \mathbb{E}_{y_{\leq 2}}[\|\mathbf{M}_2\|_2^2] &= 0.92, \\ \sum_{i=1}^2 \text{Var}(\mathbf{w}_2) &= 1. \end{aligned} \quad (24)$$

For Figure 2b, we have

$$\begin{aligned} \mathbb{E}_{y_{\leq 2}}[\|\mathbf{M}_2\|_2^2] &= 1, \\ \sum_{i=1}^2 \text{Var}(\mathbf{w}_2) &= 1. \end{aligned} \quad (25)$$

Comparing equation (24) with equation (25), we can observe that the second order moment of the martingale does not necessarily reach its maximum when the sum of the variances of $\mathbf{w}_i$ is maximized. Furthermore, it is intuitively expected that Figure 2b exhibits greater randomness than Figure 2a. Nevertheless, the above calculation results show that the sum of the variances of $\mathbf{w}_i$ cannot reflect this discrepancy, whereas the second order moment of the martingale can. This suggests that, in contrast to the sum of the variances of $\mathbf{w}_i$, the second order moment of the martingale serves as a more effective measure of output stability.

Experiments

Experimental Setup

We conduct white-box experiments (with full logit access) to empirically verify the theoretical properties of our martingale formulation. Computations are performed on an Ubuntu 22.04 system equipped with two NVIDIA RTX PRO 6000 GPUs, a 44-core Intel Xeon Platinum 8470Q CPU, 220GB RAM, Python 3.12, PyTorch 2.12.1, and CUDA 13.0.

By default, we use **Llama-3.2-3B-Instruct**. The decoding process uses pure stochastic sampling ($T = 1.0$, top- $p = 1.0$, top- $k = 0$) with no repetition penalties or external constraints. The experiments are organized into four parts: (1) verification of the first-order martingale property, (2) verification of the derived second-order moment identity, (3) scaling and generality analysis, and (4) comparison with existing uncertainty measures.

First-Order Martingale Verification

According to Theorem 1 and equation (7), the cumulative prediction error process $\mathbf{M}_t$ forms a martingale with zero first-order moment.

Therefore, we empirically examine whether the estimated cumulative prediction error remains centered around zero during autoregressive generation.

To empirically verify this property, we perform a single stochastic generation sequence with Llama-3.2-3B-Instruct. At each decoding step, the complete vocabulary-level error vector is computed. Then, $K = 300$ vocabulary dimensions are randomly sampled, and the empirical mean sequence is estimated as $\hat{\mu}_t = \frac{1}{K} \sum_{j=1}^K M_{t,v_j}$.

If the martingale property holds, this sequence should remain centered around zero without systematic drift.

Figure 3 shows the evolution of the empirical first-order moment. Although individual vocabulary dimensions exhibit random fluctuations, their averaged sequence remains close to zero throughout the generation process, which agrees with the theoretical expectation.

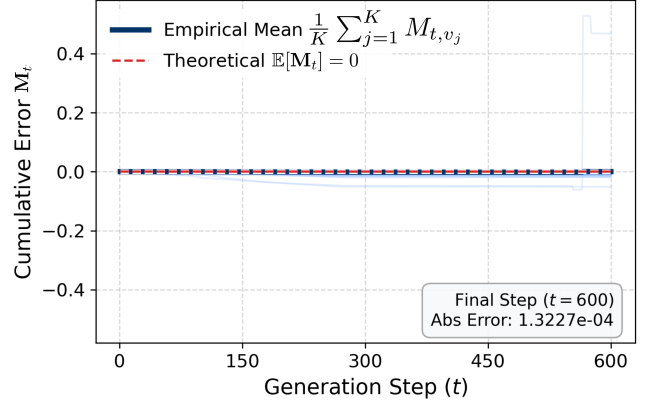


Figure 3: Empirical verification of the first-order martingale property on Llama-3.2-3B-Instruct. The sampled vocabulary dimensions exhibit random fluctuations, while their averaged cumulative error remains around the theoretical expectation $\mathbb{E}[\mathbf{M}_t] = 0$.

Second-Order Moment Identity Verification

Beyond the first-order property, Theorem 5 and equation (11) provide a tractable estimation of the second-order moment.

This result allows us to verify the theoretical quadratic variation using only the probabilities of the actually generated tokens.

To verify this relationship, we conduct Monte Carlo sampling with $K = 30$ independent generation sequences using Llama-3.2-3B-Instruct. For each sequence, we compute the empirical quadratic variation $\|\mathbf{M}_t\|_2^2$ and compare its Monte Carlo estimation with the theoretical expectation given by equation (11).

Figure 4 demonstrates that the empirical sequence closely follows the theoretical curve. The consistency between both quantities verifies the derived quadratic variation identity.

Scaling and Generality Analysis

To ensure these second-order behaviors generalize, we evaluate the Llama-3.x and Qwen2.5 families (1B to 72B parameters). Final-step statistics are summarized in Table 1, while Figure 5 presents complete trajectory comparisons. Across varying model scales, sampling temperatures, and prompt categories, the empirical moments consistently match theoretical predictions. This confirms that our derived identity characterizes an intrinsic property of autoregressive decoding rather than a model-specific artifact. Interestingly, under identical prompts, we observe that the averaged quadratic variation varies consistently across different models, hinting at its potential utility as a model-specific probabilistic fingerprint.

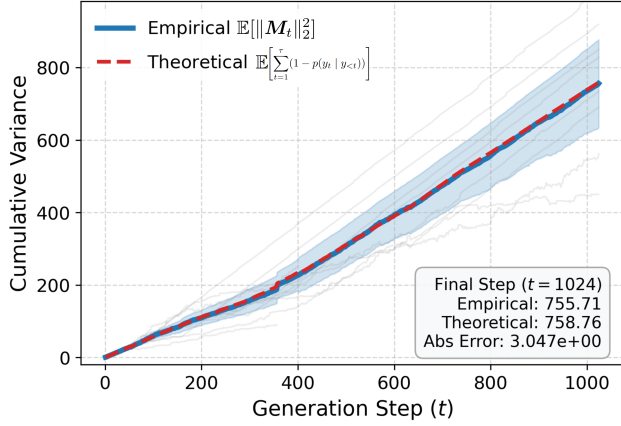


Figure 4: Verification of the second-order martingale identity on Llama-3.2-3B-Instruct. The empirical second-order moment closely matches the theoretical prediction throughout the generation sequence.

Model	Empirical	Theoretical	Error
Llama-1B	861.60	865.23	3.63
Llama-3B	796.40	781.61	14.79
Llama-8B	708.74	698.66	10.09
Llama-70B	475.45	462.09	13.36
Qwen-1.5B	609.81	622.64	12.84
Qwen-3B	434.98	436.80	1.83
Qwen-7B	282.13	283.84	1.72
Qwen-72B	210.35	214.05	3.70

Table 1: Cross-model verification of the second-order martingale identity. The empirical second-order moment consistently agrees with the theoretical quadratic variation across different model scales.

Relationship with Existing Uncertainty Measures

To interpret our derived second-order quantity, we compute its Pearson correlation against standard uncertainty proxies (predictive entropy, maximum probability, and perplexity) over 300 unconstrained trajectories from the RAID benchmark (Dugan et al. 2024).

As Table 2 shows, the length-normalized second-order quantity (NSM) is highly correlated with conventional token-level metrics. Crucially, the non-normalized cumulative form (SM) maintains only a moderate correlation. This demonstrates that our proposed martingale metric captures distinct trajectory-level global drift and cumulative stochastic dynamics that are otherwise smoothed out by conventional step-averaged uncertainty measures.

The results show that the martingale second-order moment shares related information with existing uncertainty measures. Interestingly, after normalization by sequence length, the second-order quantity becomes highly correlated with token-level uncertainty metrics. This indicates that the cumulative form preserves additional sequence-level information beyond conventional averaged uncertainty measures.

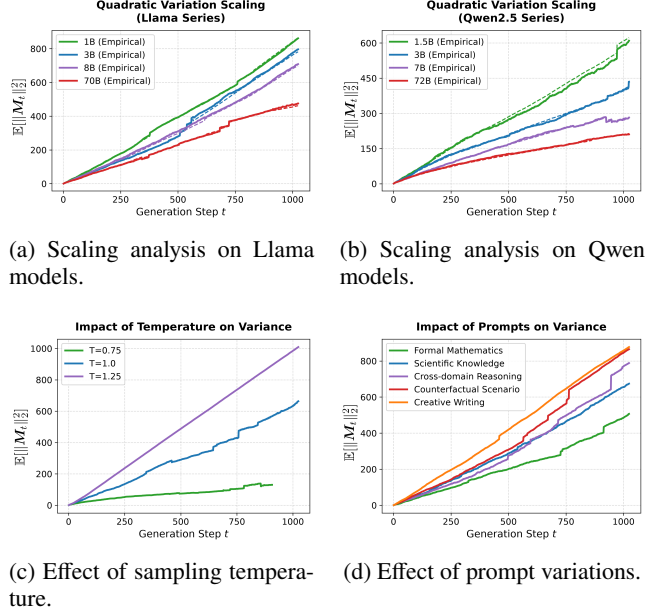


Figure 5: Generality analysis of the second-order martingale behavior. The empirical second-order moments consistently agree with the theoretical quadratic variation across different model scales, sampling temperatures, and prompt categories.

	Entropy	PPL	MaxProb
SM	0.7203	0.7057	0.7018
NSM	0.9889	0.9482	0.9950

Table 2: Pearson correlation between the martingale second-order moment (SM), its normalized form (NSM), and existing uncertainty measures.

Conclusion and Discussion

In this work, we characterize autoregressive LLM decoding as a rigorous martingale process. By defining per-step prediction errors as the difference between sampled indicators and conditional distributions, we prove that the cumulative error forms a zero-mean martingale and derive a tractable, closed-form identity for its predictable quadratic variation. Extensive white-box experiments confirm that empirical trajectories precisely track our theoretical predictions, proving these stochastic dynamics are intrinsic to autoregressive generation.

While this paper primarily focuses on theoretical characterizations, our framework offers a robust mathematical foundation for interpreting model decoding dynamics. Future work will explore leveraging these martingale properties for trajectory-level uncertainty quantification, model identification, and adaptive decoding strategies.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Aflalo, L.; et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
 Azuma, K. 1967. Weighted Sums of Certain Dependent

- Random Variables. *Tohoku Mathematical Journal*, 19(3): 357–367.
- Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.
- Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 1–122.
- Desai, S.; and Durrett, G. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 295–302.
- Doob, J. L. 1953. *Stochastic Processes*. Wiley.
- Dugan, L.; Hwang, A.; Trhlík, F.; Zhu, A.; Ludan, J. M.; Xu, H.; Ippolito, D.; and Callison-Burch, C. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12463–12492. Association for Computational Linguistics.
- Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 889–898.
- Freedman, D. A. 1975. On tail probabilities for martingales. *The Annals of Probability*, 100–118.
- Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059. PMLR.
- Garg, S.; Tsipras, D.; Liang, P.; and Valiant, G. 2022. What can transformers learn in-context? A case study of linear models. In *Advances in Neural Information Processing Systems*, volume 35, 30420–30433.
- Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 1321–1330.
- Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; et al. 2022. Training Compute-Optimal Large Language Models. In *36th Conference on Neural Information Processing Systems, NeurIPS 2022*.
- Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Zettlemoyer, L. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.
- Jiang, Z.; Araki, J.; Ding, H.; and Neubig, G. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *MIT Press - Journals*.
- Kadavath, S.; et al. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.
- Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.
- Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*.
- Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.
- Leviathan, Y.; Kalman, M.; and Matias, Y. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, 19274–19286. PMLR.
- Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2023. Contrastive Decoding: Open-ended Text Generation as Optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12286–12312.
- Li, X. L.; Thickstun, J.; Gulrajani, I.; Liang, P.; and Hashimoto, T. 2022. Diffusion-LM Improves Controllable Text Generation. In *Advances in Neural Information Processing Systems*, volume 35, 4328–4343.
- Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*.
- McDiarmid, C. 1989. On the method of bounded differences. In *Surveys in combinatorics 1989*, 148–188. Cambridge University Press.
- Meister, C.; Cotterell, R.; and Vieira, T. 2020. If Beam Search is the Answer, What is the Question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 7953–7960.
- Meister, C.; Pimentel, T.; Wiher, G.; and Cotterell, R. 2022. Typical Decoding for Natural Language Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3184–3203.
- Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 17359–17372.
- Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning*, 24950–24962. PMLR.
- Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Liao, B.; Askell, A.; et al. 2022. In-context Learning and Induction Heads. *Transformer Circuits Thread*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; et al. 2023. Llama: Open and Efficient Foundation Language Models. In *International Conference on Machine Learning*.

Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2023. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *International Conference on Learning Representations*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zhai, B.; Bendersky, M.; Bowman, S.; et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; and Weston, J. 2020. Neural Text Generation With Unlikelihood Training. In *International Conference on Learning Representations*.

A Proof

Theorem 2. *The stochastic process $\{\mathbf{M}_t\}_{t \geq 1}$ is a martingale with respect to the filtration $\mathbb{F}$.*

Proof. We have

$$\begin{aligned} & \mathbb{E}_{y_t}[\mathbf{M}_t \mid y_{<t}] \\ &= \mathbb{E}_{y_t}[\mathbf{M}_{t-1} \mid y_{<t}] + \mathbb{E}_{y_t}[\mathbf{e}_t \mid y_{<t}] \\ &= \mathbf{M}_{t-1} + \mathbb{E}_{y_t}[\mathbf{w}_t - \mathbf{g}_t \mid y_{<t}] \\ &= \mathbf{M}_{t-1} + \mathbb{E}_{y_t}[\mathbf{w}_t \mid y_{<t}] - \mathbf{g}_t \\ &= \mathbf{M}_{t-1} + \mathbf{g}_t - \mathbf{g}_t \\ &= \mathbf{M}_{t-1} \end{aligned} \quad (26)$$

$$\begin{aligned} \mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2] &\leq \mathbb{E}_{y_{\leq t}}\left[\sum_{i=1}^t \|\mathbf{e}_i\|_2\right] \leq \\ \mathbb{E}_{y_{\leq t}}\left[\sum_{i=1}^t (\|\mathbf{w}_i\|_2 + \|\mathbf{g}_i\|_2)\right] &\leq \left[\sum_{i=1}^t 2\right] = 2t < \infty \end{aligned} \quad (27)$$

$\mathbf{g}_t$ and $\mathbf{w}_t$ is decided given $y_{\leq t}$, and so is $\mathbf{e}_t$. Thus $\mathbf{M}_t$ is adapted to filtration $\mathbb{F}$. According to the definition, $\mathbf{M}_t$ is a martingale. $\square$

Theorem 3. *The second-order moment of martingale $\{\mathbf{M}_t\}_{t \geq 1}$ can be expressed as follows:*

$$\mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] = \mathbb{E}_{y_{<t}}\left[\sum_{i=1}^t (1 - \|\mathbf{g}_i\|_2^2)\right] \quad (28)$$

Proof.

$$\mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] = \mathbb{E}_{y_{<t}}[\mathbb{E}_{y_t}[\|\mathbf{M}_t\|_2^2 \mid y_{<t}]] \quad (29)$$

$$\begin{aligned} & \mathbb{E}_{y_t}[\|\mathbf{M}_t\|_2^2 \mid y_{<t}] = \mathbb{E}_{y_t}[\|\mathbf{e}_t + \mathbf{M}_{t-1}\|_2^2 \mid y_{<t}] \\ &= \mathbb{E}_{y_t}[\|\mathbf{e}_t\|_2^2 \mid y_{<t}] + \mathbb{E}_{y_t}[2\mathbf{M}_{t-1} \cdot \mathbf{e}_t \mid y_{<t}] + \\ & \quad \mathbb{E}_{y_t}[\|\mathbf{M}_{t-1}\|_2^2 \mid y_{<t}] \\ &= \mathbb{E}_{y_t}[\|\mathbf{e}_t\|_2^2 \mid y_{<t}] + 2\mathbf{M}_{t-1} \cdot \mathbb{E}_{y_t}[\mathbf{e}_t \mid y_{<t}] + \\ & \quad \|\mathbf{M}_{t-1}\|_2^2 \\ &= \mathbb{E}_{y_t}[\|\mathbf{e}_t\|_2^2 \mid y_{<t}] + \|\mathbf{M}_{t-1}\|_2^2 \end{aligned} \quad (30)$$

$$\begin{aligned} & \mathbb{E}_{y_t}[\|\mathbf{e}_t\|_2^2 \mid y_{<t}] = \mathbb{E}_{y_t}[\|\mathbf{w}_t - \mathbf{g}_t\|_2^2 \mid y_{<t}] \\ &= \mathbb{E}_{y_t}[\|\mathbf{w}_t\|_2^2 \mid y_{<t}] - \mathbb{E}_{y_t}[2\mathbf{g}_t \cdot \mathbf{w}_t \mid y_{<t}] + \\ & \quad \mathbb{E}_{y_t}[\|\mathbf{g}_t\|_2^2 \mid y_{<t}] \\ &= 1 - 2\mathbf{g}_t \cdot \mathbb{E}_{y_t}[\mathbf{w}_t \mid y_{<t}] + \|\mathbf{g}_t\|_2^2 \\ &= 1 - 2\mathbf{g}_t \cdot \mathbf{g}_t + \|\mathbf{g}_t\|_2^2 \\ &= 1 - \|\mathbf{g}_t\|_2^2 \end{aligned} \quad (31)$$

By combining equations (29), (30) and (31), we can get

$$\begin{aligned} \mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] &= \mathbb{E}_{y_{<t}}[1 - \|\mathbf{g}_t\|_2^2 + \|\mathbf{M}_{t-1}\|_2^2] = \\ & \quad \mathbb{E}_{y_{<t}}[1 - \|\mathbf{g}_t\|_2^2] + \mathbb{E}_{y_{\leq t-1}}[\|\mathbf{M}_{t-1}\|_2^2] \end{aligned} \quad (32)$$

Then the following recurrence relation holds:

$$\mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] - \mathbb{E}_{y_{\leq t-1}}[\|\mathbf{M}_{t-1}\|_2^2] = \mathbb{E}_{y_{<t}}[1 - \|\mathbf{g}_t\|_2^2] \quad (33)$$

Thus

$$\begin{aligned} & \mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] - \mathbb{E}_{y_{\leq 1}}[\|\mathbf{M}_1\|_2^2] \\ &= \mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] - \mathbb{E}_{y_1}[\|\mathbf{e}_1\|_2^2 \mid y_{<1}] \\ &= \mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] - (1 - \|\mathbf{g}_1\|_2^2) \\ &= \mathbb{E}_{y_{<t}}\left[\sum_{i=2}^t (1 - \|\mathbf{g}_i\|_2^2)\right] \end{aligned} \quad (34)$$

So

$$\mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] = \mathbb{E}_{y_{<t}}\left[\sum_{i=1}^t (1 - \|\mathbf{g}_i\|_2^2)\right] \quad (35)$$

$\square$

Theorem 4. *The second order moment of martingale $\{\mathbf{M}_t\}_{t \geq 1}$ with stopping time can be expressed as follows:*

$$\mathbb{E}_{y_{\leq l_{\max}}}[\|\mathbf{M}_\tau\|_2^2] = \mathbb{E}_{y_{\leq l_{\max}}}\left[\sum_{t=1}^{\tau} (1 - \|\mathbf{g}_t\|_2^2)\right] \quad (36)$$

Proof.

$$\begin{aligned} & \mathbb{E}_{y_{l_{\max}}}[\|\mathbf{M}_\tau\|_2^2 \mid y_{<l_{\max}}] \\ &= \mathbb{E}_{y_{l_{\max}}}\left[\sum_{t=1}^{l_{\max}} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2 \mid y_{<l_{\max}}\right] \\ &= \mathbb{E}_{y_{l_{\max}}}[\mathbb{I}_{\{\tau=l_{\max}\}} \|\mathbf{M}_{l_{\max}}\|_2^2 \mid y_{<l_{\max}}] + \\ & \quad \sum_{t=1}^{l_{\max}-1} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2 \\ &= \mathbb{I}_{\{\tau=l_{\max}\}} \mathbb{E}_{y_{l_{\max}}}[\|\mathbf{M}_{l_{\max}}\|_2^2 \mid y_{<l_{\max}}] + \\ & \quad \sum_{t=1}^{l_{\max}-1} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2 \\ &= \mathbb{I}_{\{\tau=l_{\max}\}} (1 - \|\mathbf{g}_{l_{\max}}\|_2^2 + \|\mathbf{M}_{l_{\max}-1}\|_2^2) + \\ & \quad \sum_{t=1}^{l_{\max}-1} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2 \end{aligned} \quad (37)$$

$$= \mathbb{I}_{\{\tau=l_{\max}\}}(1 - \|\mathbf{g}_{l_{\max}}\|_2^2) + \mathbb{I}_{\{\tau \geq l_{\max}-1\}} \|\mathbf{M}_{l_{\max}-1}\|_2^2 + \sum_{t=1}^{l_{\max}-2} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2$$

Taking conditional expectation on both sides with respect to $y_{<l_{\max}-1}$:

$$\begin{aligned} & \mathbb{E}_{y_{l_{\max}-1} y_{l_{\max}}} [\|\mathbf{M}_\tau\|_2^2 \mid y_{<l_{\max}-1}] \\ &= \mathbb{E}_{y_{l_{\max}-1}} [\mathbb{E}_{y_{l_{\max}}} [\|\mathbf{M}_\tau\|_2^2 \mid y_{<l_{\max}}] \mid y_{<l_{\max}-1}] \\ &= \mathbb{E}_{y_{l_{\max}-1}} [\mathbb{I}_{\{\tau=l_{\max}\}}(1 - \|\mathbf{g}_{l_{\max}}\|_2^2) \mid y_{<l_{\max}-1}] + \\ & \quad \mathbb{I}_{\{\tau \geq l_{\max}-1\}} \mathbb{E}_{y_{l_{\max}-1}} [\|\mathbf{M}_{l_{\max}-1}\|_2^2 \mid y_{<l_{\max}-1}] + \\ & \quad \sum_{t=1}^{l_{\max}-2} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2 \\ &= \mathbb{E}_{y_{l_{\max}-1}} [\mathbb{I}_{\{\tau=l_{\max}\}}(1 - \|\mathbf{g}_{l_{\max}}\|_2^2) \mid y_{<l_{\max}-1}] + \\ & \quad \mathbb{I}_{\{\tau \geq l_{\max}-1\}}(1 - \|\mathbf{g}_{l_{\max}-1}\|_2^2 + \|\mathbf{M}_{l_{\max}-2}\|_2^2) + \\ & \quad \sum_{t=1}^{l_{\max}-2} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2 \\ &= \mathbb{E}_{y_{l_{\max}-1}} [\mathbb{I}_{\{\tau \geq l_{\max}\}}(1 - \|\mathbf{g}_{l_{\max}}\|_2^2) \mid y_{<l_{\max}-1}] + \\ & \quad \mathbb{I}_{\{\tau \geq l_{\max}-1\}}(1 - \|\mathbf{g}_{l_{\max}-1}\|_2^2) + \\ & \quad \mathbb{I}_{\{\tau \geq l_{\max}-2\}} \|\mathbf{M}_{l_{\max}-2}\|_2^2 + \sum_{t=1}^{l_{\max}-3} \mathbb{I}_{\{\tau=t\}} \|\mathbf{M}_t\|_2^2 \end{aligned} \quad (38)$$

Then, taking conditional expectations successively with respect to $y_{<i}$ ($i = l_{\max} - 2, \dots, 1$), we have

$$\begin{aligned} & \mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2] \\ &= \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{l_{\max}} \mathbb{I}_{\{\tau \geq t\}} (1 - \|\mathbf{g}_t\|_2^2) \right] \\ &= \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} (1 - \|\mathbf{g}_t\|_2^2) \right] \end{aligned} \quad (39)$$

□

Theorem 5. *The second-order moment of martingale $\{\mathbf{M}_t\}_{t \geq 1}$ can be expressed as follows:*

$$\mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2] = \mathbb{E}_{y_{\leq t}} \left[\sum_{i=1}^t (1 - p(y_i \mid y_{<i})) \right] \quad (40)$$

Proof.

$$\begin{aligned} & \mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2] \\ &= \mathbb{E}_{y_{<t}} \left[\sum_{i=1}^t (1 - \|\mathbf{g}_i\|_2^2) \right] \\ &= \sum_{i=1}^t \mathbb{E}_{y_{<t}} [1 - \|\mathbf{g}_i\|_2^2] \\ &= \sum_{i=1}^t \mathbb{E}_{y_{<i}} [1 - \|\mathbf{g}_i\|_2^2] \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^t \sum_{y_{<i} \in \mathcal{V}^{i-1}} p(y_{<i}) (1 - \|\mathbf{g}_i\|_2^2) \\ &= \sum_{i=1}^t \sum_{y_{<i} \in \mathcal{V}^{i-1}} p(y_{<i}) \left(1 - \sum_{y_i \in \mathcal{V}} p^2(y_i \mid y_{<i}) \right) \\ &= \sum_{i=1}^t \sum_{y_{<i} \in \mathcal{V}^{i-1}} (p(y_{<i}) - \sum_{y_i \in \mathcal{V}} p(y_{<i}) p^2(y_i \mid y_{<i})) \quad (41) \\ &= \sum_{i=1}^t \sum_{y_{<i} \in \mathcal{V}^{i-1}} (p(y_{<i}) - \sum_{y_i \in \mathcal{V}} p(y_{\leq i}) p(y_i \mid y_{<i})) \\ &= \sum_{i=1}^t \sum_{y_{\leq i} \in \mathcal{V}^i} p(y_{\leq i}) (1 - p(y_i \mid y_{<i})) \\ &= \sum_{i=1}^t \mathbb{E}_{y_{\leq i}} [1 - p(y_i \mid y_{<i})] \\ &= \sum_{i=1}^t \mathbb{E}_{y_{\leq t}} [1 - p(y_i \mid y_{<i})] \\ &= \mathbb{E}_{y_{\leq t}} \left[\sum_{i=1}^t (1 - p(y_i \mid y_{<i})) \right] \end{aligned}$$

□

Theorem 6. *The second order moment of martingale $\{\mathbf{M}_t\}_{t \geq 1}$ with stopping time can be expressed as follows:*

$$\mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2] = \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} (1 - p(y_t \mid y_{<t})) \right] \quad (42)$$

Proof.

$$\begin{aligned} & \mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2] = \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{l_{\max}} \mathbb{I}_{\{\tau \geq t\}} (1 - \|\mathbf{g}_t\|_2^2) \right] \\ &= \sum_{t=1}^{l_{\max}} \mathbb{E}_{y_{\leq l_{\max}}} [\mathbb{I}_{\{\tau \geq t\}} (1 - \|\mathbf{g}_t\|_2^2)] \\ &= \sum_{t=1}^{l_{\max}} \mathbb{E}_{y_{<t}} [\mathbb{I}_{\{\tau \geq t\}} (1 - \|\mathbf{g}_t\|_2^2)] \\ &= \sum_{t=1}^{l_{\max}} \sum_{y_{<t} \in \mathcal{V}^{t-1}} p(y_{<t}) \mathbb{I}_{\{\tau \geq t\}} \left(1 - \sum_{y_t \in \mathcal{V}} p^2(y_t \mid y_{<t}) \right) \\ &= \sum_{t=1}^{l_{\max}} \sum_{y_{<t} \in \mathcal{V}^{t-1}} \mathbb{I}_{\{\tau \geq t\}} (p(y_{<t}) - \sum_{y_t \in \mathcal{V}} p(y_{<t}) p^2(y_t \mid y_{<t})) \\ &= \sum_{t=1}^{l_{\max}} \sum_{y_{<t} \in \mathcal{V}^{t-1}} \mathbb{I}_{\{\tau \geq t\}} (p(y_{<t}) - \sum_{y_t \in \mathcal{V}} p(y_{\leq t}) p(y_t \mid y_{<t})) \\ &= \sum_{t=1}^{l_{\max}} \sum_{y_{\leq t} \in \mathcal{V}^t} p(y_{\leq t}) \mathbb{I}_{\{\tau \geq t\}} (1 - p(y_t \mid y_{<t})) \quad (43) \end{aligned}$$

$$\begin{aligned}
&= \sum_{t=1}^{l_{\max}} \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} (1 - p(y_t | y_{<t}))] \\
&= \sum_{t=1}^{l_{\max}} \mathbb{E}_{y_{\leq l_{\max}}} [\mathbb{I}_{\{\tau \geq t\}} (1 - p(y_t | y_{<t}))] \\
&= \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{l_{\max}} \mathbb{I}_{\{\tau \geq t\}} (1 - p(y_t | y_{<t})) \right] \\
&= \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} (1 - p(y_t | y_{<t})) \right]
\end{aligned}$$

□

Theorem 7. *The second-order moment of martingale (with stopping time) can be expressed as follows:*

$$\mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2] = \mathbb{E}_{y_{\leq t}} \left[\sum_{i=1}^t \mathbf{e}_i^2 \right] \quad (44)$$

$$\mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2] = \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} \mathbf{e}_t^2 \right] \quad (45)$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{y_{\leq l_{\max}}} [\mathbf{e}_i^2] = \mathbb{E}_{y_{\leq i}} [\mathbf{e}_i^2] \\
&= \mathbb{E}_{y_{\leq i}} [(\mathbf{w}_i - \mathbf{g}_i)^2] \\
&= \mathbb{E}_{y_{\leq i}} [\mathbf{w}_i^2 - 2\mathbf{w}_i \cdot \mathbf{g}_i + \mathbf{g}_i^2] \\
&= \mathbb{E}_{y_{\leq i}} [\mathbf{w}_i^2] - 2\mathbb{E}_{y_{\leq i}} [\mathbf{w}_i \cdot \mathbf{g}_i] + \mathbb{E}_{y_{\leq i}} [\mathbf{g}_i^2] \\
&= 1 - 2\mathbb{E}_{y_{\leq i}} [\mathbb{E}_{y_i} [\mathbf{w}_i \cdot \mathbf{g}_i | y_{<i}]] + \mathbb{E}_{y_{\leq i}} [\mathbf{g}_i^2] \\
&= 1 - 2\mathbb{E}_{y_{\leq i}} [\mathbf{g}_i^2] + \mathbb{E}_{y_{\leq i}} [\mathbf{g}_i^2] \\
&= \mathbb{E}_{y_{\leq i}} [1 - \mathbf{g}_i^2] \\
&\mathbb{E}_{y_{\leq l_{\max}}} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{e}_t^2] = \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{e}_t^2] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} (\mathbf{w}_t - \mathbf{g}_t)^2] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} (\mathbf{w}_t^2 - 2\mathbf{w}_t \cdot \mathbf{g}_t + \mathbf{g}_t^2)] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{E}_{y_t} [\mathbb{I}_{\{\tau \geq t\}} (1 - 2\mathbf{w}_t \cdot \mathbf{g}_t + \mathbf{g}_t^2) | y_{<t}]] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} \mathbb{E}_{y_t} [1 - 2\mathbf{w}_t \cdot \mathbf{g}_t + \mathbf{g}_t^2 | y_{<t}]] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} (1 - \mathbf{g}_t^2)]
\end{aligned} \quad (46)$$

$$\begin{aligned}
&\mathbb{E}_{y_{\leq l_{\max}}} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{e}_t^2] = \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{e}_t^2] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} (\mathbf{w}_t - \mathbf{g}_t)^2] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} (\mathbf{w}_t^2 - 2\mathbf{w}_t \cdot \mathbf{g}_t + \mathbf{g}_t^2)] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{E}_{y_t} [\mathbb{I}_{\{\tau \geq t\}} (1 - 2\mathbf{w}_t \cdot \mathbf{g}_t + \mathbf{g}_t^2) | y_{<t}]] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} \mathbb{E}_{y_t} [1 - 2\mathbf{w}_t \cdot \mathbf{g}_t + \mathbf{g}_t^2 | y_{<t}]] \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} (1 - \mathbf{g}_t^2)]
\end{aligned} \quad (47)$$

By combining equations (46) and (47) with equations (28) and (36), respectively, we obtain equations (44) and (45). □

Theorem 8. *For the second order moments $\mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2]$ and $\mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2]$, the following relation holds:*

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) = \mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2] + \sum_{i=1}^t \text{Var}(\mathbf{g}_i) \quad (48)$$

$$\begin{aligned}
&\mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} \text{Var}(\mathbf{w}_t) \right] \\
&= \mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2] + \mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} \text{Var}(\mathbf{g}_t) \right]
\end{aligned} \quad (49)$$

Then

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) \geq \mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2] \quad (50)$$

$$\mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} \text{Var}(\mathbf{w}_t) \right] \geq \mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2] \quad (51)$$

Proof. By the law of total variance, we have

$$\begin{aligned}
&\text{Var}(\mathbf{w}_i) \\
&= \mathbb{E}_{y_{<i}} [\text{Var}(\mathbf{w}_i | y_{<i})] + \text{Var}(\mathbb{E}_{y_i} [\mathbf{w}_i | y_{<i}]) \\
&= \mathbb{E}_{y_{<i}} [\mathbb{E}_{y_i} [(\mathbf{w}_i - \mathbb{E}_{y_i} [\mathbf{w}_i | y_{<i}])^2 | y_{<i}]] + \\
&\quad \text{Var}(\mathbb{E}_{y_i} [\mathbf{w}_i | y_{<i}]) \\
&= \mathbb{E}_{y_{<i}} [\mathbb{E}_{y_i} [(\mathbf{w}_i - \mathbf{g}_i)^2 | y_{<i}]] + \text{Var}(\mathbf{g}_i) \\
&= \mathbb{E}_{y_{\leq i}} [\mathbf{e}_i^2] + \text{Var}(\mathbf{g}_i)
\end{aligned} \quad (52)$$

$$\begin{aligned}
&\text{Var}(\mathbb{I}_{\{\tau \geq t\}} \mathbf{w}_t) = \mathbb{E}_{y_{<t}} [\text{Var}(\mathbb{I}_{\{\tau \geq t\}} \mathbf{w}_t | y_{<t})] + \\
&\quad \text{Var}(\mathbb{E}_{y_t} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{w}_t | y_{<t}]) \\
&= \mathbb{E}_{y_{<t}} [\mathbb{E}_{y_t} [(\mathbb{I}_{\{\tau \geq t\}} \mathbf{w}_t - \mathbb{E}_{y_t} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{w}_t | y_{<t}])^2 | y_{<t}]] + \\
&\quad \text{Var}(\mathbb{E}_{y_t} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{w}_t | y_{<t}]) \\
&= \mathbb{E}_{y_{<t}} [\mathbb{E}_{y_t} [(\mathbb{I}_{\{\tau \geq t\}} \mathbf{w}_t - \mathbb{I}_{\{\tau \geq t\}} \mathbb{E}_{y_t} [\mathbf{w}_t | y_{<t}])^2 | y_{<t}]] + \\
&\quad \text{Var}(\mathbb{I}_{\{\tau \geq t\}} \mathbb{E}_{y_t} [\mathbf{w}_t | y_{<t}]) \\
&= \mathbb{E}_{y_{<t}} [\mathbb{E}_{y_t} [\mathbb{I}_{\{\tau \geq t\}} (\mathbf{w}_t - \mathbf{g}_t)^2 | y_{<t}]] + \text{Var}(\mathbb{I}_{\{\tau \geq t\}} \mathbf{g}_t) \\
&= \mathbb{E}_{y_{\leq t}} [\mathbb{I}_{\{\tau \geq t\}} \mathbf{e}_t^2] + \text{Var}(\mathbb{I}_{\{\tau \geq t\}} \mathbf{g}_t)
\end{aligned} \quad (53)$$

Combining equations (52) and (53) with equations (44) and (45), respectively, yields equations (48) and (49).

Noting that

$$\sum_{i=1}^t \text{Var}(\mathbf{g}_i) \geq 0 \quad (54)$$

$$\mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} \text{Var}(\mathbf{g}_t) \right] \geq 0 \quad (55)$$

from equations (48) and (49) we obtain equations (50) and (51). □

Theorem 9. *The following equivalences hold:*

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) = 0 \iff \mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2] = 0 \quad (56)$$

$$\begin{aligned}
&\mathbb{E}_{y_{\leq l_{\max}}} \left[\sum_{t=1}^{\tau} \text{Var}(\mathbf{w}_t) \right] = 0 \iff \\
&\quad \mathbb{E}_{y_{\leq l_{\max}}} [\|\mathbf{M}_\tau\|_2^2] = 0
\end{aligned} \quad (57)$$

Proof. From equations (50) and (51), the left-to-right implications of equations (56) and (57) are obvious. We first consider the right-to-left implication of equation (56).

From equation (44), $\mathbb{E}_{y_{\leq t}} [\|\mathbf{M}_t\|_2^2] = 0$ is equivalent to $\forall 1 \leq i \leq t : \mathbb{E}_{y_{\leq t}} [\mathbf{e}_i^2] = 0$. This implies that $\forall 1 \leq i \leq t : (\mathbf{w}_i = \mathbf{g}_i \text{ a.e.})$, meaning that for output sequence $y_1 y_2 \dots, y_t$

with nonzero probability, $\mathbf{g}_i$ is a one-hot vector. That is, there is only one output sequence with probability 1, and all other output sequence have probability 0. This means that the output at each token is fixed, and $\sum_{i=1}^t \text{Var}(\mathbf{w}_i) = 0$.

For equation (57), we consider those t with $P(\tau = t) > 0$; the rest of the proof is similar to that of equation (56). $\square$

Theorem 10. *The following equivalences hold:*

$$\sum_{i=1}^t \text{Var}(\mathbf{w}_i) = t(1 - \frac{1}{|\mathcal{V}|}) \iff \quad (58)$$

$$\forall 1 \leq i \leq t : \mathbb{E}_{y_{<i}}[\mathbf{g}_i] = \frac{1}{|\mathcal{V}|} \mathbf{1}$$

$$\mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] = t(1 - \frac{1}{|\mathcal{V}|}) \iff \quad (59)$$

$$\forall 1 \leq i \leq t : (\mathbf{g}_i = \frac{1}{|\mathcal{V}|} \mathbf{1} \text{ a.e.})$$

where $\mathbf{1}$ is the all-one vector.

The following implication hold:

$$\begin{aligned} \mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] = t(1 - \frac{1}{|\mathcal{V}|}) \Rightarrow \\ \sum_{i=1}^t \text{Var}(\mathbf{w}_i) = t(1 - \frac{1}{|\mathcal{V}|}) \end{aligned} \quad (60)$$

Proof.

$$\begin{aligned} & \sum_{i=1}^t \text{Var}(\mathbf{w}_i) \\ &= \sum_{i=1}^t (\mathbb{E}_{y_{\leq i}}[\mathbf{w}_i^2] - \mathbb{E}_{y_{\leq i}}[\mathbf{w}_i]^2) \\ &= \sum_{i=1}^t (1 - (\mathbb{E}_{y_{<i}}[\mathbb{E}_{y_i}[\mathbf{w}_i | y_{<i}]]))^2 \quad (61) \\ &= \sum_{i=1}^t (1 - \mathbb{E}_{y_{<i}}^2[\mathbf{g}_i]) \\ &\leq \sum_{i=1}^t (1 - \frac{1}{|\mathcal{V}|}) = t(1 - \frac{1}{|\mathcal{V}|}) \end{aligned}$$

Equality holds in equation (61) if and only if $\forall 1 \leq i \leq t : \mathbb{E}_{y_{<i}}[\mathbf{g}_i] = \frac{1}{|\mathcal{V}|} \mathbf{1}$.

Considering equation (28), we have

$$\begin{aligned} & \mathbb{E}_{y_{\leq t}}[\|\mathbf{M}_t\|_2^2] \\ &= \mathbb{E}_{y_{<t}} \left[\sum_{i=1}^t (1 - \|\mathbf{g}_i\|_2^2) \right] \\ &= \sum_{i=1}^t (1 - \mathbb{E}_{y_{<i}}[\|\mathbf{g}_i\|_2^2]) \quad (62) \\ &= \sum_{i=1}^t (1 - \mathbb{E}_{y_{<i}}[\|\mathbf{g}_i\|_2^2]) \end{aligned}$$

$$\begin{aligned} & \leq \sum_{i=1}^t (1 - \frac{1}{|\mathcal{V}|}) \\ &= t(1 - \frac{1}{|\mathcal{V}|}) \end{aligned}$$

Equality holds in equation (62) if and only if $\forall 1 \leq i \leq t : \mathbb{E}_{y_{<i}}[\|\mathbf{g}_i\|_2^2] = \frac{1}{|\mathcal{V}|}$, i.e., $\forall 1 \leq i \leq t : (\mathbf{g}_i = \frac{1}{|\mathcal{V}|} \mathbf{1} \text{ a.e.})$. Since $\forall 1 \leq i \leq t : (\mathbf{g}_i = \frac{1}{|\mathcal{V}|} \mathbf{1} \text{ a.e.}) \Rightarrow \forall 1 \leq i \leq t : \mathbb{E}_{y_{<i}}[\mathbf{g}_i] = \frac{1}{|\mathcal{V}|} \mathbf{1}$, equation (60) holds. $\square$