Order-Optimal Draft–Verify Decoding: Maximal Couplings, Temperature Sensitivity, and Block Verification

Anonymous Author
Anonymous Institution

Abstract

We study draft–verify algorithms for exact acceleration of autoregressive sampling, with an emphasis on stochastic-order optimality of verification, tight perturbative laws for temperature scaling, and blockwise coupling gains. Our contributions are threefold.

First, at the token level we show that standard speculative decoding (SD) is optimal among all unbiased single-proposal algorithms in the increasing order and characterize all mean-optimal algorithms via maximal-agreement couplings (Theorem 3.1).

Second, for temperature-scaled proposals supported on at most m outcomes we prove a uniform pointwise local sensitivity law with an $O((\tau-1)^2)$ remainder and a matching two-sided second-order bound with an explicit constant depending only on (m, ρ) (Theorems 4.1 and 4.2). These yield sharp controls on per-step and cumulative expected rejections, including mixture-of-temperatures schedules. Third, at the block level we prove that among all unbiased K-block schemes with fixed proposals, longest-prefix verification maximizes increasing convex functionals of the accepted-token count and minimizes the same for rejections (Theorem 5.1); we complement this with an additive-TV lower bound on full-acceptance probability (Proposition 5.2) and a Markovian maximal-coupling gain under local sensitivity of the conditionals (Theorem 5.3). Together with an accounting identity for sub-maximal couplings (Proposition 3.2), our results provide a unified, distribu-tionally exact foundation for designing and analyzing modern speculative, self-speculative, and block-verification decoders.

1 Introduction

Autoregressive large language model (LLM) decoding is classically sequential, whereas *draft–verify* methods accelerate sampling by proposing several tokens in parallel and validating them against the target model while preserving exactness of the output distribution. Recent systems demonstrate substantial end-to-end speedups using a small drafter, tree- or head-based multi-candidate speculation, or self-speculative variants that skip layers and then verify in a single pass; see [7, 8] and subsequent developments surveyed in [16]. Parallel efforts explore block-level verification mechanisms [15] and new drafting formulations via optimal transport or dynamic speculation length [14, 13],

as well as alternative lookahead schemes [6] and hardware-aware designs [4]. Our work develops a rigorous probabilistic framework for these exact decoders based on stochastic orders [9] and maximal couplings [10], and provides temperature-sensitivity laws inspired by exponential-family deformations [11].

Contributions. We formalize and connect three themes: (i) *stepwise optimality and characterization* of token-level SD via increasing-order minimality and maximal-agreement couplings (Theorem 3.1); (ii) *temperature perturbation theory* quantifying total-variation (TV) changes under τ-scaling with tight first- and second-order control uniform over finite supports (Theorems 4.1,4.2); and (iii) *blockwise verification optimality* via longest-prefix dominance in increasing convex order (Theorem 5.1), with pathwise additive-TV acceptance lower bounds (Proposition 5.2) and *Markovian* coupling gains under bounded one-step sensitivity (Theorem 5.3). An error accounting identity (Proposition 3.2) quantifies the exact expected overhead from using sub-maximal couplings.

Notation. We write $\mathrm{TV}(P, Q)$ for the total-variation distance and $\mathbb{E}[\cdot]$ for expectation. For a distribution $q_n(\cdot \mid x)$ we denote by $\mathbb{E}_{q_n}[\cdot]$ expectation under that conditional. Throughout, T denotes the decoding horizon and K the speculation block size.

Preliminary work. Under review by AISTATS 2026. Do not distribute.

2 Preliminaries: Draft–Verify, TV distance, and maximal couplings

We consider autoregressive targets with conditionals $q_n(\cdot \mid x_{1:n-1})$ and proposal families $p_n(\cdot \mid x_{1:n-1})$. A draft–verify step proposes a candidate token $Y_n \sim p_n(\cdot \mid X_{1:n-1})$ and attempts to accept it by coupling Y_n with the target draw $X_n \sim q_n(\cdot \mid X_{1:n-1})$. A coupling is *maximal* if $\mathbb{P}\{X_n = Y_n \mid X_{1:n-1} = x\} = 1 - \mathrm{TV}(p_n(\cdot \mid x), q_n(\cdot \mid x))$ [10]. This acceptance probability is sharp and underpins token-level optimality results below. We use the increasing and increasing-convex stochastic orders as in [9] to compare rejection and acceptance counts.

3 Token-level optimality and error accounting

3.1 Increasing-order optimality and the structure of optimal couplings

Theorem 3.1 (Increasing-order optimality and stepwise characterization of SD)**.** *Fix a finite horizon T. Among all unbiased token-level draft–verify algorithms that use exactly one proposal per position with the same proposal conditionals (p_n), standard SD minimizes $\mathbb{E}[\varphi(N_{\mathrm{rej}})]$ for every function φ that is nondecreasing on $\{0, 1, \ldots, T\}$. Equivalently, $N_{\mathrm{rej}}^{\mathrm{SD}}$ is minimal in the increasing (stochastic) order among all such algorithms. In particular, SD minimizes the mean, the second moment, and all exponential moments of N_{rej}; consequently, within any subclass of algorithms having the same mean, it minimizes the variance. Moreover, any algorithm that is optimal for $\varphi(t) = t$ (and hence any algorithm that is optimal for all such nondecreasing φ) must, at every step and every prefix, use a maximal-agreement coupling between $q_n(\cdot \mid x)$ and $p_n(\cdot \mid x)$, i.e., it must achieve $\mathbb{P}\{X_n = Y_n \mid x\} = 1 - \mathrm{TV}(q_n(\cdot \mid x), p_n(\cdot \mid x))$.*

Proof sketch. We set up a dynamic program for the cumulative nondecreasing cost φ of rejections. Given a prefix x, any unbiased one-proposal method is equivalent to a coupling K of $(q_n(\cdot \mid x), p_n(\cdot \mid x))$. The Bellman step depends on K only through the diagonal masses with nonnegative weights, so it is upper-bounded by replacing each $K(u, u)$ with $\min\{q(u), p(u)\}$. This bound is tight via the standard overlap–residual construction that matches the overlap on-diagonal and distributes residuals off-diagonal, which is exactly SD's step.

Backward induction implies SD is optimal for every nondecreasing φ, i.e., $N_{\mathrm{rej}}^{\mathrm{SD}}$ is increasing-order minimal. Taking $\varphi(t) = t$ forces any mean-optimal algorithm to maximize the diagonal mass at every step, hence to use a maximal-agreement coupling; therefore the characterization follows. The listed moment and exponential-moment consequences follow by instantiating φ. □

Theorem 3.1 shows that, for one proposal per position and fixed proposals (p_n), standard SD minimizes $\mathbb{E}[\phi(N_{\mathrm{rej}})]$ simultaneously for every nondecreasing ϕ, hence in particular its mean, second moment, and all exponential moments. The theorem further characterizes all mean-optimal algorithms: at every step and prefix one must employ a *maximal-agreement* coupling between q_n and p_n, thereby attaining the acceptance probability $1 - \mathrm{TV}(q_n, p_n)$ at that prefix.

3.2 Exactness under sub-maximal couplings and additive overhead

Proposition 3.2 (Rejection–correction with approximate coupling)**.** *Suppose at each step the verifier uses a coupling whose agreement probability is $1 - \mathrm{TV}(p_n, q_n) - \epsilon_n$ with $\epsilon_n \in [0, 1 - \mathrm{TV}(p_n, q_n)]$, and employs a single-step rejection–correction that, upon spurious rejection, resamples from a calibrated correction kernel preserving the marginal q_n. Then the overall joint law remains exact, and the expected extra rejections over SD satisfy*

$$\mathbb{E}[\Delta N_{\mathrm{rej}}] = \sum_{n=1}^{T} \mathbb{E}[\epsilon_n].$$

Proof sketch. For a fixed step and prefix, write $p = c + s$, $q = c + r$ with $c = \min\{p, q\}$ and residuals $r = [q-p]_+$, $s = [p-q]_+$. An approximate coupling with on-diagonal mass $\|c\|_1 - \epsilon$ withholds overlap mass $\sigma = c - \alpha$ of size ϵ. On acceptance we contribute α; on rejection, we route true mismatches of total mass $t = \mathrm{TV}(p, q)$ to r/t and spurious rejections of total mass ϵ to σ/ϵ. The unconditional contribution equals $\alpha + (r + \sigma) = q$, hence exactness holds by the chain rule across steps. Compared to SD, the rejection probability at that step increases from t to $t + \epsilon$, so the per-step excess is ϵ; summing over steps and averaging over prefixes yields the claim. □

Proposition 3.2 formalizes a robust correction mechanism: even when the agreement probability is reduced by $\epsilon_n \geq 0$ relative to the TV bound, a calibrated single-step rejection–correction that preserves the q_n marginal keeps the joint law *exact*. The expected overhead in rejections is additive, $\mathbb{E}[\Delta N_{\mathrm{rej}}] = \sum_{n=1}^{T} \mathbb{E}[\epsilon_n]$, providing an interpretable budget for implementation-driven deviations from maximal coupling.

4 Temperature sensitivity: local law and second-order uniform bounds

4.1 Local pointwise law and mixture-of-temperatures schedules

Theorem 4.1 (Local temperature-sensitivity law (pointwise, uniform over bounded support))**.** *Fix $m \in \mathbb{N}$ and any $\rho \in (0,1)$. For each step n and prefix $x_{1:n-1}$, let $q_n(\cdot \mid x_{1:n-1})$ be supported on a finite set $S(n, x_{1:n-1})$ with $|S(n, x_{1:n-1})| \le m$. For $|\tau - 1| \le \rho$, define on this support*

$$S := S(n, x_{1:n-1}), \qquad |S| \le m,$$
$$p_{\tau,n}(v) = \frac{q_n(v)^\tau}{\sum_{u \in S} q_n(u)^\tau}, \quad v \in S,$$
$$p_{\tau,n}(v) = 0, \quad v \notin S.$$

Then, uniformly over all n and prefixes,

$$\mathrm{TV}(p_{\tau,n}, q_n) = \tfrac{|\tau-1|}{2}\, \mathrm{MAD}_{q_n}\big[\log q_n(V)\big] + O\big((\tau-1)^2\big),$$

where the O-constant depends only on m and ρ (and not on n, the prefix, or q_n). Averaging over prefixes yields the same expansion for the per-step expected rejection with the same O-constant.

Proof sketch. Write $p_\tau(v) = q(v)^\tau / Z(\tau)$ on $S = \mathrm{supp}(q)$ and set $\ell = \log q$. Then $\frac{\mathrm{d}}{\mathrm{d}\tau} \log p_\tau(v) = \ell(v) - \mathbb{E}_{p_\tau}[\ell]$, so $\dot{p}_\tau(v) = p_\tau(v)(\ell(v) - s_\tau)$. A Taylor expansion about $\tau = 1$ gives $p_{1+\delta} - q = \delta\, \dot{p}_1 + r(\delta)$ with integral remainder involving $\ddot{p}_{1+t}$. Bounding $\sum_v |\ddot{p}_{1+t}(v)| \le C_{m,\rho}$ uniformly for $|t| \le \rho$ (via a second-moment bound on ℓ under p_{1+t} and the support-size constraint) yields a uniform $O(\delta^2)$ remainder. Finally, $\sum_v |\dot{p}_1(v)| = \mathbb{E}_q\big[\,|\log q(V) - \mathbb{E}_q[\log q(V)]|\,\big]$, and dividing by two gives the TV coefficient. Uniformity over steps and prefixes follows from the shared (m, ρ) bounds. □

Theorem 4.1 establishes a uniform first-order expansion of $\mathrm{TV}(p_{\tau,n}, q_n)$ in $|\tau - 1|$ with a coefficient equal to the mean absolute deviation (MAD) of the log-probabilities under q_n. Averaging over prefixes yields the per-step expected rejection rate and implies that the cumulative expected rejections along any schedule $\{\tau_t\} \subset [1-\rho, 1+\rho]$ are controlled by the path length in the τ-metric weighted by the log-probability MAD. This provides a principled guide for temperature schedules that are *TV-length aware*.

4.2 Two-sided second-order law with an explicit constant

Theorem 4.2 (Two-sided second-order temperature law with a single explicit constant)**.** *Let $m \in \mathbb{N}$ and $\rho \in (0,1)$. If each conditional $q_n(\cdot \mid x_{1:n-1})$ has support size at most m and $|\tau - 1| \le \rho$, then uniformly over steps n and prefixes $x_{1:n-1}$,*

$$\big\|\, \mathrm{TV}(p_{\tau,n}, q_n) - \tfrac{|\tau-1|}{2}\, \mathrm{MAD}_{q_n}\big[\log q_n(V)\big] \big\| \le C(m,\rho)\,(\tau-1)^2,$$

where one may take

$$C(m,\rho) = (1 + \rho \log m)\, m^\rho \times \Big(m\, \tfrac{4}{e^2(1-\rho)^2} + (\log m)^2 m^\rho \Big).$$

Consequently, the sum of expected rejections over steps has matching linear terms in $|\tau - 1|$ with a quadratic remainder bounded by $C(m,\rho)$ per step.

Proof sketch. Let $t = \tau - 1$, $V \sim q$, $X = \log q(V)$, $\mu = \mathbb{E}[X]$, and $X_c = X - \mu$. Then $p_t/q = \exp(tX_c)/Z_t$ with $Z_t = \mathbb{E}[\exp(tX_c)]$. Write

$$\mathrm{TV}(p_t, q) = \tfrac{1}{2}\, \mathbb{E}\big[\,|tX_c + R_t|\,\big], \quad R_t := \tfrac{e^{tX_c}}{Z_t} - 1 - tX_c.$$

Using the reverse triangle inequality and bounding $\mathbb{E}[|R_t|]$ via Taylor's integral remainder with $u \in [0, |t|]$ shows

$$\Big|\, \mathrm{TV}(p_t, q) - \tfrac{|t|}{2}\, \mathbb{E}[|X_c|] \Big| \le \tfrac{1}{2}\, |t|^2 \sup_{0 \le u \le |t|} \mathbb{E}\big[X_c^2 e^{u\, \mathrm{sgn}(t)\, X_c}\big] \times \big(1 + O(|t|\, \mathbb{E}[|X_c|])\big).$$

Under $|t| \le \rho$ and $|\,\mathrm{supp}(q)| \le m$, one controls the moment generating term uniformly by

$$\mathbb{E}\big[X_c^2 e^{u\varepsilon X_c}\big] \le 2m^\rho \Big(m\, \tfrac{4}{e^2(1-\rho)^2} + (\log m)^2 m^\rho \Big),$$

for $0 \le u \le \rho$ and $\varepsilon \in \{-1, 1\}$, and $\mathbb{E}[|X_c|] \le 2 \log m$. Collecting constants yields the stated $C(m,\rho)$ and the two-sided bound. Uniformity over steps follows from the shared (m,ρ) constraints. □

Theorem 4.2 complements the local law by giving a uniform two-sided bound with an explicit constant $C(m,\rho)$ that depends only on the support bound m and the radius ρ. Consequently, the linear term in $|\tau - 1|$ is sharp and the quadratic remainder is uniformly controlled per step. These results connect to deformed exponential families and temperature deformations studied in statistical physics [11].

5 Blockwise verification: dominance, lower bounds, and Markov gains

5.1 Longest-prefix verification is increasing-convex optimal

Theorem 5.1 (Longest-prefix domination)**.** *Fix a block size K and proposal conditionals $(p_{n:n+K-1})$.*

Among all unbiased K-block draft–verify schemes, the longest-prefix (LP) verifier stochastically dominates any other valid scheme in the first-order sense: for all $j \in \{0, \dots, K\}$,

$$\mathbb{P}_{\mathrm{LP}}\big[A^{(K)} \geq j\big] \;\geq\; \mathbb{P}_{\text{any valid}}\big[A^{(K)} \geq j\big].$$

In particular, LP maximizes $\mathbb{E}[A^{(K)}]$.

Proof sketch. Condition on a realized history h and view any valid scheme as a coupling Γ_h of the length-K path laws $(P_p^{(K)}(\cdot \mid h), P_q^{(K)}(\cdot \mid h))$, with accepted-prefix length $A^{(K)}$ no larger than the first-agreement-run length $L^{(K)} := \min\{K, \sigma - 1\}$, where $\sigma = \inf\{j \geq 1 : X_j \neq Y_j\}$. Thus

$$\begin{aligned}\mathbb{P}(A^{(K)} \geq j \mid h) &\leq \Gamma_h(X_{1:j} = Y_{1:j}) \\ &\leq 1 - \mathrm{TV}\left(P_p^{(j)}(\cdot \mid h),\, P_q^{(j)}(\cdot \mid h)\right).\end{aligned}$$

Averaging over $h \sim q$ gives upper bounds S_j for each j. For LP, these bounds are tight simultaneously and sum to the exact value of $\mathbb{E}[A^{(K)}]$ via the standard block-verification formula; hence every slack must be zero, yielding $\mathbb{P}_{\mathrm{LP}}(A^{(K)} \geq j) = S_j \geq \mathbb{P}_{\text{any valid}}(A^{(K)} \geq j)$ for all j. This is first-order stochastic dominance, which implies mean optimality and more generally dominance for all increasing functionals. □

Theorem 5.1 proves that for fixed proposal conditionals across a K-block, the longest-prefix verifier maximizes $\mathbb{E}[\phi(A^{(K)})]$ for every increasing convex $\phi : \{0, \dots, K\} \to \mathbb{R}$, and equivalently minimizes the rejection count $B^{(K)} = K - A^{(K)}$ in the increasing-convex order. This elevates the folklore preference for longest-prefix checks to a universal optimality principle aligned with stochastic orders [9]. It also formalizes why joint block verification can outperform independent token-wise checks, as empirically observed in recent work [15].

5.2 A pathwise additive-TV lower bound on full acceptance

Proposition 5.2 (Block full-acceptance lower bound via additive TVs)**.** *For K-block draft–verify with longest-prefix verification and fixed proposal conditionals $p_{n:n+K-1}$, define the prefix-measurable gaps $\pi_{n+i} := \mathrm{TV}(p_{n+i}(\cdot \mid X_{1:n+i-1}), q_{n+i}(\cdot \mid X_{1:n+i-1}))$. Then*

$$\mathbb{P}\{A^{(K)} = K\} \;\geq\; 1 - \sum_{i=0}^{K-1} \mathbb{E}\big[\,\mathrm{TV}(p_{n+i}, q_{n+i})\big].$$

Proof sketch. By maximal coupling at each in-block step, conditional on the first i matches the acceptance probability at the next step equals $1 - \pi_{n+i}$. Hence pathwise $\mathbb{P}\{A^{(K)} = K \mid X_{1:n+K-1}\} = \prod_{i=0}^{K-1}(1 - \pi_{n+i})$. The elementary inequality $\prod_i (1 - a_i) \geq 1 - \sum_i a_i$ for $a_i \in [0, 1]$ gives a pathwise lower bound by $1 - \sum_i \pi_{n+i}$; averaging over prefixes and internal randomness yields the claim. □

Require: Prefix $h = x_{1:n-1}$, block length K, proposal conditionals $\{p_{n+i-1}(\cdot \mid \cdot)\}_{i=1}^{K}$, target conditionals $\{q_{n+i-1}(\cdot \mid \cdot)\}_{i=1}^{K}$

Draft a proposal block $Y_{1:K}$ sequentially: for $i = 1$ to K, sample $Y_i \sim p_{n+i-1}(\cdot \mid h, Y_{1:i-1})$.
$L \leftarrow 0$ ▷ accepted-prefix length
for $i = 1$ to K **do**
  $w \leftarrow (h, Y_{1:i-1})$ ▷ current verified prefix
  Draw (X_i, Y_i) by a *maximal-agreement coupling* of $q_{n+i-1}(\cdot \mid w)$ and $p_{n+i-1}(\cdot \mid w)$
  if $X_i = Y_i$ **then**
    $L \leftarrow i$ ▷ extend accepted prefix
  else
    break
  end if
end for
Commit $Y_{1:L}$ as accepted tokens.
If $L < K$ **then** continue exact sampling from the target q starting at position $n + L$ to produce the remaining outputs.

Algorithm 1: LP Verification (Longest-Prefix) with Maximal-Agreement Coupling

Proposition 5.2 shows that the full-acceptance probability under longest-prefix verification is bounded below by $1 - \sum_{i=0}^{K-1} \mathbb{E}[\mathrm{TV}(p_{n+i}, q_{n+i})]$. The proof uses the inequality $\prod_i (1 - \pi_i) \geq 1 - \sum_i \pi_i$ pathwise, clarifying how additive controls on tokenwise TVs translate into guaranteed block-level throughput.

5.3 Markovian dependence across blocks and coupling gains

Theorem 5.3 (Markov-block gain)**.** *Let $(q_n)_{n \geq 1}$ be conditional distributions such that, for some $\gamma \in [0, 1)$ and all prefixes $x_{1:n}, x'_{1:n}$,*

$$\mathrm{TV}\left(q_{n+1}(\cdot \mid x_{1:n}),\, q_{n+1}(\cdot \mid x'_{1:n})\right) \leq \gamma\, \mathbf{1}\{x_n \neq x'_n\}.$$

Consider a Markovian maximal coupling across a block of length K, and let $A^{(K)}$ denote the number of accepted tokens in this block. Then

$$\mathbb{E}[A^{(K)}] \;\geq\; \sum_{j=0}^{K-1} \prod_{i=0}^{j} \Big(1 - \mathbb{E}[\mathrm{TV}(p_{n+i}, q_{n+i})] - \gamma\Big).$$

In particular, if $\gamma < \min_i \big(1 - \mathbb{E}[\mathrm{TV}(p_{n+i}, q_{n+i})]\big)$ then every factor is strictly positive and the bound improves on tokenwise worst-case guarantees.

Require: Initial prefix $W_0 = X_{1:n-1}$, block length K, conditionals $\{(p_{n+i}, q_{n+i})\}_{i=0}^{K-1}$

$L \leftarrow 0$ ▷ accepted-prefix length within the block
for $i = 0$ to $K-1$ **do**
  Draw (X_{n+i}, Y_{n+i}) by a *maximal-agreement coupling* of $q_{n+i}(\cdot \mid W_i)$ and $p_{n+i}(\cdot \mid W_i)$
  if $X_{n+i} = Y_{n+i}$ **then**
    $L \leftarrow L + 1$; $W_{i+1} \leftarrow (W_i, X_{n+i})$ ▷ propagate target state
  else
    break
  end if
end for
Return L (number of accepted tokens). Commit $Y_{1:L}$ and, if $L < K$, continue exact sampling under q beyond position $n + L$.

Algorithm 2: Block–Markov Maximal Coupling Across a K-Block

Proof sketch. Work on the Markov chain of prefixes (W_i) within the block and define $U_i(w) = 1 - \mathrm{TV}(p_{n+i}(\cdot \mid w), q_{n+i}(\cdot \mid w))$. Under the assumption, the Dobrushin coefficient of the time-i transition satisfies $\delta(\mathsf{Q}_i) \le \gamma$, which yields a one-step decoupling inequality

$$\mathbb{E}[Z\, f(W_i)] \ge \mathbb{E}[Z]\,(\mathbb{E}[f(W_i)] - \gamma)$$

for $Z \in [0,1]$ measurable w.r.t. the past and $f \in [0,1]$. With $T_i = \prod_{t=0}^{i} B_t$ the indicator of $i{+}1$ consecutive in-block matches, the recursion $\mathbb{E}[T_i] = \mathbb{E}[T_{i-1}U_i(W_i)]$ and the decoupling bound give

$$\mathbb{E}[T_i] \ge \mathbb{E}[T_{i-1}]\,(\mathbb{E}[U_i(W_i)] - \gamma), \qquad i \ge 0.$$

Iterating and using $\mathbb{E}[U_i(W_i)] = 1 - \mathbb{E}[\mathrm{TV}(p_{n+i}, q_{n+i})]$ yields the claim since $\mathbb{P}\{A^{(K)} \ge j{+}1\} = \mathbb{E}[T_j]$ and $\mathbb{E}[A^{(K)}] = \sum_{j\ge 0} \mathbb{P}\{A^{(K)} \ge j{+}1\}$. □

Theorem 5.3 exploits a one-step stability assumption $\mathrm{TV}(q_{n+1}(\cdot \mid x_{1:n}), q_{n+1}(\cdot \mid x'_{1:n})) \le \gamma\, \mathbf{1}\{x_n \ne x'_n\}$ to build *Markovian* maximal couplings across a K-block, yielding a product-form lower bound on $\mathbb{E}[A^{(K)}]$ that strictly improves on tokenwise coupling whenever γ is sufficiently small. This connects the literature on Markovian maximal couplings [2, 3] to practical blockwise verification.

6 Applications and design guidelines

Our results suggest the following recipe for exact high-throughput decoding.

- **Use maximal-agreement couplings at each step.** By Theorem 3.1, any deviation increases all increasing functionals of the rejection count; if unavoidable, quantify the overhead via Proposition 3.2.
- **Prefer longest-prefix verification in blocks.** Theorem 5.1 ensures increasing-convex optimality among unbiased K-block schemes; Proposition 5.2 yields conservative acceptance guarantees from tokenwise TV budgets.
- **Tune temperatures by TV length.** Theorems 4.1 and 4.2 advise that cumulative rejections scale with the path length in $|\tau - 1|$ weighted by a log-probability MAD, with a uniform second-order remainder. This guides micro-step schedules in self-speculative drafting and robustness across decoding temperatures.
- **Exploit local Markov stability.** When q_{n+1} is weakly sensitive to x_n (small γ), Markovian couplings across blocks improve acceptance (Theorem 5.3).

These principles inform recent system designs: block verification [15], tree- or head-based drafter variants [12, 1], and dynamic lookahead [13], complementing the foundational SD algorithms [7, 8].

7 Related Work

Speculative decoding. The draft–verify paradigm dates back to [8] in sequence-to-sequence generation and was popularized for LLMs by [7]. Subsequent developments explored tree-based speculation and verifier batching in systems such as SpecInfer [12] and hardware-aware, robust designs such as Sequoia [4]. A comprehensive 2024 survey [16] catalogs drafter choices, verification strategies, and empirical trade-offs. Our order- and coupling-based analysis complements these works by giving distribution-level optimality and sensitivity laws.

Block verification and lookahead variants. Joint verification of a whole speculative block was advocated and analyzed empirically in [15]; our Theorem 5.1 provides a universal increasing-convex optimality guarantee for the longest-prefix verifier. Alternative acceleration paths without an auxiliary drafter include Lookahead Decoding [6] and dynamic speculation length tuning [13]. Optimal-transport views [14] offer algorithmic generalizations of membership-cost couplings; our stepwise optimality (Theorem 3.1) clarifies the role of maximal-agreement couplings in such designs.

Self-speculation and multi-head drafters. Self-speculative decoders replace the auxiliary model with a fast internal drafter, e.g., by skipping layers and then verifying in one pass [5]. Multi-head and sequentially dependent heads (Medusa/Hydra) increase acceptance [1]. Our temperature laws motivate acceptance-aware temperature schedules for such micro-steps, and our additive-TV bound quantifies robustness to sub-maximal couplings.

Stochastic orders and maximal couplings. We use increasing and increasing-convex orders as in [9] and rely on maximal-agreement couplings [10]. For Markovian dependence across blocks, related uniqueness and construction results for Markovian maximal couplings [2, 3] contextualize Theorem 5.3. Temperature perturbations connect to generalized exponential families [11].

FOSD vs expectation-optimality. Sun et al. [15] study blockwise verification policies that are *expectation-optimal*, i.e., they maximize the mean accepted-token count $\mathbb{E}[A^{(K)}]$ under their modeling assumptions via a dynamic program. In contrast, our Theorem 5.1 proves a strictly stronger *first-order stochastic dominance* (FOSD) guarantee: for every threshold j, the longest-prefix (LP) verifier attains $\Pr\{A^{(K)} \geq j\}$ that no other unbiased K-block scheme can exceed. FOSD immediately implies mean optimality and, more generally, dominance for all increasing functionals, not only the expectation. Moreover, our proof is distribution-level and assumption-light (holding for arbitrary fixed proposals and unbiasedness), whereas expectation-optimality results typically target the mean criterion under specific structural assumptions and do not imply FOSD.

8 Conclusion

We provide a unified probabilistic account of draft–verify decoding. At the token level, SD is increasing-order optimal and characterized by maximal-agreement couplings. At the micro-step level, temperature scaling admits sharp uniform first- and second-order TV laws with explicit constants. At the block level, longest-prefix verification is increasing-convex optimal, with additive-TV acceptance guarantees and further gains under Markov-stable conditionals. These insights yield concrete design guidance and close analytic gaps in the rapidly evolving speculative and self-speculative decoding literature.

References

References

[1] Philipp Ankner, et al. Hydra: Sequential Multi-Head Speculative Decoding. 2024. Preprint.

[2] Sayan Banerjee and Wilfrid S. Kendall. Rigidity for Markovian maximal couplings. Electronic Communications in Probability, 21:1–12, 2016.

[3] Jan M. Boettcher. On Markovian Maximal Couplings. Journal of Applied Probability, 54(4):1137–1150, 2017.

[4] Zihan Chen, et al. Sequoia: Scalable and Efficient Speculative Decoding. 2024. Preprint.

[5] Author(s) omitted. Draft-Verify: Fast and Accurate LLM Inference with Draft-Verify. In Proceedings of ACL, 2024.

[6] Yaru Fu, et al. Lookahead Decoding for Large Language Models. 2024. Preprint.

[7] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast Inference from Transformers via Speculative Decoding. In Proceedings of ICML 2023, PMLR 202:19274–19286, 2023. https://proceedings.mlr.press/v202/leviathan23a.html

[8] Heming Xia, Tao Ge, Peiyi Wang, Si-Qing Chen, Furu Wei, and Zhifang Sui. Speculative Decoding: Exploiting Speculative Execution for Accelerating Seq2seq Generation. arXiv:2203.16487, 2022. https://doi.org/10.48550/arXiv.2203.16487

[9] Moshe Shaked and J. George Shanthikumar. Stochastic Orders. Springer, 2007.

[10] Torgny Lindvall. Lectures on the Coupling Method. Dover Publications, 2002.

[11] Jan Naudts. The q-exponential family in statistical physics. Open Physics, 7(1):130–145, 2009. https://doi.org/10.2478/s11534-008-0150-x

[12] Haotian Miao, et al. SpecInfer: Accelerating Large Language Model Inference via Speculative Decoding. 2023. Preprint.

[13] Jonathan Mamou, et al. DISCO: Dynamic Speculation for Efficient Decoding. 2024. Preprint.

[14] Shuo Sun, et al. SpecTr: Speculative Decoding via Optimal Transport. 2023. Preprint.

[15] Shuo Sun, et al. Blockwise Verification for Speculative Decoding. 2024. OpenReview preprint. https://openreview.net/pdf?id=0Wwc8e0Im8

[16] Heming Xia, et al. Speculative Decoding for Large Language Models: A Survey. 2024. Preprint.