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

Anonymous

September 22, 2025

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, distributionally 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.

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 $\{0, 1, \dots, T\}$. 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. Fix $T \in \mathbb{N}$ and a finite alphabet $\mathcal{V}$. Let q be the target autoregressive (AR) model with conditionals $q_n(\cdot \mid x_{1:n-1})$ and let p denote the draft conditionals $p_n(\cdot \mid x_{1:n-1})$. Consider any unbiased token-level draft–verify algorithm A that uses exactly one proposal per position and the given (p_n); unbiasedness means the output law is q. At step n, after the accepted prefix $X_{1:n-1} = x_{1:n-1}$, the algorithm draws $Y_n \sim p_n(\cdot \mid x_{1:n-1})$ and outputs $X_n \in \mathcal{V}$ with marginal $q_n(\cdot \mid x_{1:n-1})$. Define $R_n \equiv \mathbf{1}\{X_n \neq Y_n\}$ and $N_{\mathrm{rej}} \equiv \sum_{n=1}^{T} R_n$.

Step 1 (Reduction to couplings at each step). For each n and prefix $x \equiv x_{1:n-1}$, the pair (X_n, Y_n) under A induces a coupling $K_n^A(\cdot, \cdot \mid x)$ of $q_n(\cdot \mid x)$ and $p_n(\cdot \mid x)$. Conversely, any family of couplings $K_n(\cdot, \cdot \mid x)$ can be realized by first sampling $Y_n \sim p_n(\cdot \mid x)$ and then $X_n \sim K_n(\cdot \mid Y_n, x)$, which preserves $X_n \sim q_n(\cdot \mid x)$. Since unbiasedness enforces $X_{1:n-1} \sim q$, expectations at step n are taken with respect to $x \sim q$ for every unbiased algorithm.

Step 2 (Dynamic program). Let φ be nondecreasing on $\{0, 1, \ldots, T\}$. For $n \in \{1, \ldots, T+1\}$, $x \in \mathcal{V}^{n-1}$ and $s \in \{0, 1, \ldots, n-1\}$ define

$$W_n(x, s) \equiv \inf_{\text{valid one-proposal algorithms from step } n} \mathbb{E}\big[\, \varphi(s + \textstyle\sum_{j=n}^{T} R_j) \,\big|\, X_{1:n-1} = x\big].$$

Set $W_{T+1}(x, s) = \varphi(s)$ for all x, s. For $n \le T$ and any coupling K of $q_n(\cdot \mid x)$ and $p_n(\cdot \mid x)$ we have

$$\mathbb{E}_{(u,v)\sim K}\big[\, W_{n+1}(xu,\ s + \mathbf{1}\{u \neq v\})\,\big] = \mathbb{E}_{u\sim q_n(\cdot|x)}\big[W_{n+1}(xu, s)\big] + \mathbb{E}_{(u,v)\sim K}\big[\Delta_{n+1}(xu; s)\, \mathbf{1}\{u \neq v\}\big],$$

where $\Delta_{n+1}(xu; s) \equiv W_{n+1}(xu, s+1) - W_{n+1}(xu, s) \ge 0$, by backward induction from the assumption that φ is nondecreasing on $\{0, \ldots, T\}$. Hence, for fixed (x, s), minimizing the Bellman expression is equivalent to maximizing

$$\sum_{u\in\mathcal{V}} \Delta_{n+1}(xu; s)\, K(u, u) \quad \text{over all couplings } K \text{ of } q_n(\cdot \mid x) \text{ and } p_n(\cdot \mid x).$$

Step 3 (Optimal stepwise coupling via a tight upper bound and explicit attainment). Fix (x, s) and abbreviate $q \equiv q_n(\cdot \mid x)$, $p \equiv p_n(\cdot \mid x)$, and $\Delta(u) \equiv \Delta_{n+1}(xu; s) \ge 0$. For any coupling K of (q, p), each diagonal entry satisfies

$$K(u, u) \le \min\{q(u), p(u)\} \qquad (\forall u \in \mathcal{V}),$$

so, using nonnegativity of the weights,

$$\sum_u \Delta(u)\, K(u, u) \;\le\; \sum_u \Delta(u)\, \min\{q(u), p(u)\}. \tag{$*$}$$

We claim the upper bound $(*)$ is attained. Define the overlap and residuals

$$\rho(u) \equiv \min\{q(u), p(u)\}, \qquad r(u) \equiv [\, q(u) - p(u)\,]_+, \qquad s(u) \equiv [\, p(u) - q(u)\,]_+.$$

Then $\sum_u r(u) = \sum_u s(u) = \mathrm{TV}(q, p)$ and $r(u)\, s(u) = 0$ for every u. Choose any matrix L supported on $\{(i, j) : r(i) > 0,\ s(j) > 0\}$ with row sums $r(i)$ and column sums $s(j)$. Now define a coupling K^* by

$$K^*(u, u) = \rho(u) \quad (\forall u), \qquad K^*(i, j) = L(i, j) \ \text{ for } \ i \in \{r > 0\},\ j \in \{s > 0\},$$

and $K^*(i, j) = 0$ otherwise. Then K^* has marginals (q, p), achieves $K^*(u, u) = \min\{q(u), p(u)\}$ for all u, and hence attains equality in $(*)$. Consequently, for the Bellman step at (x, s), any maximal-agreement coupling (one with $K(u, u) = \min\{q(u), p(u)\}$ for all u) is optimal. In particular, this coupling agrees with the standard token-level SD rule: with probability $\sum_u \min\{q(u), p(u)\} = 1 - \mathrm{TV}(q, p)$ it sets $X = Y$, and otherwise draws X from the residual $[q - p]_+ / \mathrm{TV}(q, p)$.

Backward induction (sufficiency). Since at every (x, s) a maximal-agreement coupling is optimal for the Bellman step, applying it at each $n = 1, \ldots, T$ yields a globally optimal policy. Hence

$$\mathbb{E}_{\mathrm{SD}}[\varphi(N_{\mathrm{rej}})] = \mathbb{E}[W_1(\emptyset, 0)] \;\le\; \mathbb{E}_A[\varphi(N_{\mathrm{rej}})]$$

for every unbiased one-proposal algorithm A with proposal conditionals (p_n). Equivalently, $N_{\mathrm{rej}}^{\mathrm{SD}}$ is minimal in the increasing (stochastic) order among all such algorithms.

Necessity of maximal agreement for mean-optimality. Take $\varphi(t) = t$. Then $\Delta_{n+1}(xu; s) \equiv 1$ for all states, so the Bellman step reduces to maximizing $\sum_u K(u,u)$ subject to $K(u,u) \leq \min\{q(u), p(u)\}$, which forces $K(u,u) = \min\{q(u), p(u)\}$ for every u. Thus any algorithm that minimizes $\mathbb{E}[N_{\text{rej}}]$ must, at every step and prefix, use a maximal-agreement coupling; in particular, any algorithm that is simultaneously optimal for all nondecreasing φ coincides with SD on the agreement event and attains $\mathbb{P}\{X_n = Y_n \mid x\} = 1 - \text{TV}(q_n(\cdot \mid x), p_n(\cdot \mid x))$ at each step.

Consequences. Because the inequality holds for every nondecreasing φ on $\{0, \dots, T\}$:

- With $\varphi(t) = t$, SD minimizes $\mathbb{E}[N_{\text{rej}}]$ and attains the instance-dependent lower bound $\sum_{n=1}^{T} \mathbb{E}_{x \sim q}[\text{TV}(p_n(\cdot \mid x), q_n(\cdot \mid x))]$.
- With $\varphi(t) = t^2$, SD minimizes the second moment $\mathbb{E}[N_{\text{rej}}^2]$. Consequently, within any class of unbiased algorithms having the same mean $\mathbb{E}[N_{\text{rej}}]$, SD minimizes the variance.
- With $\varphi(t) = e^{\lambda t}$ for $\lambda > 0$, SD minimizes all exponential moments of N_{rej}.

Thus, among all unbiased token-level draft–verify algorithms using exactly one proposal per position with the same conditionals (p_n), standard SD minimizes $\mathbb{E}[\varphi(N_{\text{rej}})]$ for every nondecreasing φ on $\{0, \dots, T\}$, establishing increasing-order optimality and characterizing stepwise optimality via maximal-agreement couplings.

To summarize the main inequality concisely,

$$\mathbb{E}_{\text{SD}}\big[\varphi(N_{\text{rej}})\big] \;\leq\; \mathbb{E}_A\big[\varphi(N_{\text{rej}})\big] \quad \text{for all nondecreasing } \varphi : \{0, \dots, T\} \to \mathbb{R} \text{ and all admissible } A. \quad \square$$

Theorem 3.1 shows that, for one proposal per position and fixed proposals (p_n), standard SD minimizes $\mathbb{E}[\phi(N_{\text{rej}})]$ 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 $1 - \text{TV}(q_n, p_n)$ 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 $1 - \text{TV}(p_n, q_n) - \epsilon_n$ with $\epsilon_n \in [0, 1 - \text{TV}(p_n, q_n)]$, 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 incurred over SD satisfy*

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

Proof. Let $T < \infty$ and, for each step n and prefix $x_{1:n-1}$, let $p_n(\cdot \mid x_{1:n-1})$ and $q_n(\cdot \mid x_{1:n-1})$ be distributions on a common measurable space. Suppose the verifier uses a coupling at step n whose agreement probability equals $1 - \text{TV}(p_n, q_n) - \epsilon_n$ with $\epsilon_n \in [0, 1 - \text{TV}(p_n, q_n)]$, and upon a spurious rejection performs a single-step rejection–correction that resamples from a calibrated kernel preserving the marginal $q_n(\cdot \mid x_{1:n-1})$.

Fix a step n and a prefix $x \equiv x_{1:n-1}$. For brevity write $p := p_n(\cdot \mid x)$ and $q := q_n(\cdot \mid x)$, and set $t := \text{TV}(p, q)$. Let

$$\begin{aligned} &c \equiv p \wedge q, \qquad r \equiv [q - p]_+, \qquad s \equiv [p - q]_+, \\ &\|c\|_1 = 1 - t, \quad \|r\|_1 = \|s\|_1 = t, \quad q = c + r, \quad p = c + s. \end{aligned}$$

By assumption, the verifier employs a coupling π of (p, q) whose on-diagonal mass is $\mathbb{P}_\pi\{Y = Z\} = 1 - t - \epsilon$ for some $\epsilon \in [0, 1 - t]$. Let the on-diagonal measure be

$$\alpha(v) \equiv \pi\{Y = Z = v\}, \qquad v \in \mathcal{V}.$$

Then $0 \le \alpha \le c$ (as measures) and $\|\alpha\|_1 = 1 - t - \epsilon$. Define the spurious shortfall measure $\sigma := c - \alpha \ge 0$, so $\|\sigma\|_1 = \epsilon$.

Single-step rejection–correction. Condition on the coupling's outcome:

- Accept if $Y = Z$, and set $X_n := Y$; the unconditional contribution to $\mathcal{L}(X_n \mid x)$ is the measure α.
- Otherwise, a rejection occurs. We distinguish (measurably within the coupling) between (i) true mismatches of total mass t, and (ii) spurious rejections of total mass ϵ (the withheld common mass σ). On rejection we resample X_n from the calibrated kernel that draws

$$\begin{aligned} X_n &\sim r/t \quad \text{for true mismatches (when } t > 0\text{)}, \\ X_n &\sim \sigma/\epsilon \quad \text{for spurious rejections (when } \epsilon > 0\text{)}. \end{aligned}$$

Thus the unconditional contribution from all rejections to $\mathcal{L}(X_n \mid x)$ is

$$t \cdot \frac{r}{t} + \epsilon \cdot \frac{\sigma}{\epsilon} = r + \sigma.$$

Hence the final conditional law at step n is

$$\mathcal{L}(X_n \mid x) \;=\; \alpha + (r + \sigma) \;=\; (\alpha + \sigma) + r \;=\; c + r \;=\; q.$$

Since this holds for every prefix x, we have for all $x_{1:T}$ by the chain rule

$$\mathbb{P}\{X_{1:T} = x_{1:T}\} \;=\; \prod_{n=1}^{T} q_n(x_n \mid x_{1:n-1}) \;=\; q(x_{1:T}),$$

so the joint law is exact.

Extra rejections. In standard SD (maximal agreement), the rejection probability at step n given x equals $t = \mathrm{TV}(p, q)$. Under the approximate coupling it is $t + \epsilon$. Therefore the per-step excess rejection probability equals ϵ, and summing over steps and averaging over random prefixes yields

$$\mathbb{E}[\Delta N_{\text{rej}}] \;=\; \sum_{n=1}^{T} \mathbb{E}[\epsilon_n]. \qquad \square$$

This also covers the boundary cases: if $1 - t = 0$ then necessarily $\epsilon = 0$; if $t = 0$ (resp. $\epsilon = 0$) the corresponding rejection branch is vacuous.

Proposition 3.2 formalizes a robust correction mechanism: even when the agreement probability is reduced by $\epsilon_n \ge 0$ 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, $\mathbb{E}[\Delta N_{\text{rej}}] = \sum_{n=1}^{T} \mathbb{E}[\epsilon_n]$, 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 a distribution 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 the temperature-scaled proposal on this support by*

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

Then, uniformly over all n and prefixes,

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

where the O-constant depends only on m and ρ (and not on n, the prefix, or q_n). Consequently, averaging over prefixes yields the per-step expected rejection

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

with the same O-constant.

Proof. Fix a step n and prefix $x_{1:n-1}$. Write $q \equiv q_n(\cdot \mid x_{1:n-1})$, let $S := \mathrm{supp}(q)$, and assume $|S| \le m$. Fix any $\rho \in (0,1)$ and restrict to $|\tau - 1| \le \rho$ so that $\tau > 0$. Define, for $v \in S$,

$$p_\tau(v) = \frac{q(v)^\tau}{Z(\tau)}, \qquad Z(\tau) := \sum_{u \in S} q(u)^\tau, \quad \text{and set } p_\tau(v) = 0 \text{ for } v \notin S.$$

Let $\ell(v) := \log q(v)$ for $v \in S$ and $s_\tau := \mathbb{E}_{p_\tau}[\ell]$. Then, for $v \in S$,

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \log p_\tau(v) = \ell(v) - s_\tau =: h_\tau(v), \qquad \Rightarrow \qquad \dot{p}_\tau(v) = p_\tau(v)\, h_\tau(v).$$

Differentiating again and using $\dot{s}_\tau = \sum_{u \in S} \ell(u) \dot{p}_\tau(u) = \mathrm{Var}_{p_\tau}(\ell)$ gives, for $v \in S$,

$$\ddot{p}_\tau(v) = p_\tau(v)\big(h_\tau(v)^2 - \mathrm{Var}_{p_\tau}(\ell)\big). \tag{1}$$

For $v \notin S$ we have $p_\tau(v) \equiv 0$ for $\tau > 0$, hence $\dot{p}_\tau(v) = \ddot{p}_\tau(v) = 0$.

Taylor's formula with integral remainder about $\tau = 1$ yields, for $\delta := \tau - 1$ and all v,

$$p_{1+\delta}(v) = q(v) + \delta\, \dot{p}_1(v) + r_v(\delta), \qquad r_v(\delta) = \int_0^\delta (\delta - t)\, \ddot{p}_{1+t}(v)\, \mathrm{d}t.$$

Summing absolute values and using $\big||a+b| - |a|\big| \le |b|$ gives

$$\Big|\sum_v |p_{1+\delta}(v) - q(v)| - |\delta| \sum_v |\dot{p}_1(v)|\Big| \le \sum_v |r_v(\delta)| \le \int_0^{|\delta|} (|\delta| - t) \sum_v |\ddot{p}_{1+t}(v)|\, \mathrm{d}t. \tag{2}$$

From (1) on S and zeros off S,

$$\sum_v |\ddot{p}_\tau(v)| \le \sum_{v \in S} p_\tau(v)\big(h_\tau(v)^2 + \mathrm{Var}_{p_\tau}(\ell)\big) = 2\, \mathrm{Var}_{p_\tau}(\ell) \le 2\, \mathbb{E}_{p_\tau}[\ell^2].$$

For $|\tau-1| \le \rho < 1$, write $a := \tau \in [1-\rho, 1+\rho]$. Then

$$\mathbb{E}_{p_a}[\ell^2] = \frac{\sum_{v\in S} q(v)^a\, \ell(v)^2}{\sum_{v\in S} q(v)^a}.$$

With $q(v) = e^{-s}$ and $s \ge 0$, we have $q(v)^a\, \ell(v)^2 = e^{-as}s^2 \le \max_{s\ge 0} s^2 e^{-as} = 4/(a^2e^2) \le 4/((1-\rho)^2e^2)$. Hence

$$\sum_{v\in S} q(v)^a\, \ell(v)^2 \le \frac{4|S|}{(1-\rho)^2e^2}.$$

Moreover, for $a \in [1-\rho, 1]$ we have $\sum_{v\in S} q(v)^a \ge 1$, and for $a \in [1, 1+\rho]$, $\sum_{v\in S} q(v)^a \ge |S|^{1-a} \ge |S|^{-\rho}$. Thus for all $a \in [1-\rho, 1+\rho]$,

$$\sum_{v\in S} q(v)^a \ge |S|^{-\rho}.$$

Therefore,

$$\sup_{|\tau-1|\le\rho} \sum_v |\ddot{p}_\tau(v)| \;\le\; 2 \sup_{|\tau-1|\le\rho} \mathbb{E}_{p_\tau}[\ell^2] \;\le\; \frac{8}{e^2}\,\frac{|S|^{1+\rho}}{(1-\rho)^2} \;\le\; \frac{8}{e^2}\,\frac{m^{1+\rho}}{(1-\rho)^2} =: C_{m,\rho}. \tag{3}$$

Combining (2)–(3) yields $\sum_v |r_v(\delta)| \le (C_{m,\rho}/2)\,\delta^2$ for $|\delta| \le \rho$. At $\tau = 1$, $p_1 = q$ and, for $v \in S$, $\dot{p}_1(v) = q(v)(\ell(v) - \mathbb{E}_q[\ell])$ (while $\dot{p}_1(v) = 0$ for $v \notin S$), hence

$$\sum_v |\dot{p}_1(v)| = \sum_{v\in S} q(v)\, |\ell(v) - \mathbb{E}_q[\ell]| = \mathbb{E}_q\big[\,|\log q(V) - \mathbb{E}_q[\log q(V)]|\,\big].$$

Using $\mathrm{TV}(p,q) = \frac{1}{2}\sum_v |p-q|$, we have, uniformly for $|\tau-1| \le \rho$,

$$\mathrm{TV}(p_\tau, q) = \frac{|\tau-1|}{2}\, \mathbb{E}_q\big[\,|\log q(V) - \mathbb{E}_q[\log q(V)]|\,\big] + O\big((\tau-1)^2\big),$$

where the O-constant depends only on (m,ρ) and not on q. This establishes the asserted pointwise expansion, uniformly over all n and prefixes with $|\operatorname{supp}(q_n)| \le m$. Averaging over prefixes (by linearity of expectation) yields the stated expansion for the per-step expected rejection with the same O-constant:

$$\tau_n(\tau) = \mathbb{E}_{x_{1:n-1}}\big[\mathrm{TV}(p_{\tau,n}, q_n)\big] = \tfrac{|\tau-1|}{2}\, \mathbb{E}_{x_{1:n-1}}\Big[\mathbb{E}_{q_n}\big[\,|\log q_n(V) - \mathbb{E}_{q_n}[\log q_n(V)]|\,\big]\Big] + O\big((\tau-1)^2\big). \quad \square$$

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 $x_{1:n-1}$,*

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

where

$$C(m,\rho) = (1+\rho\log m)\, m^{\rho}\Big(m\, B_1(\rho) + (\log m)^2 m^{\rho}\Big), \qquad B_1(\rho) = \frac{4}{e^2(1-\rho)^2}.$$

Equivalently, the two-sided bounds hold with the same constant: one may take $C_-(m,\rho) = C_+(m,\rho) = C(m,\rho)$. Consequently, $\sum_n \mathbb{E}[\mathrm{TV}(p_{\tau,n}, q_n)]$ has matching linear terms in $|\tau - 1|$ with a quadratic remainder bounded by $C(m,\rho)$ per step, uniformly in n.

Proof. Fix $m \in \mathbb{N}$ and $\rho \in (0,1)$. For any step n and prefix $x_{1:n-1}$, let $q \equiv q_n(\cdot \mid x_{1:n-1})$ be supported on a set S with $|S| \le m$, and for $|\tau - 1| \le \rho$ define

$$p_\tau(v) = \frac{q(v)^\tau}{\sum_{u\in S} q(u)^\tau}, \qquad v \in S.$$

Write $t := \tau - 1$ with $|t| \le \rho$. Let $V \sim q$, set $X := \log q(V)$, $\mu := \mathbb{E}_q[X] = \sum_v q(v)\log q(v) \in [-\log m, 0]$, and $X_c := X - \mu$. Then

$$p_t(v) = \frac{q(v)^{1+t}}{\sum_u q(u)^{1+t}} = q(v)\,\frac{e^{tX_c(v)}}{Z_t}, \qquad Z_t := \mathbb{E}_q\big[e^{tX_c}\big].$$

Hence $\mathrm{TV}(p_t, q) = \frac{1}{2}\sum_v |p_t(v) - q(v)| = \frac{1}{2}\,\mathbb{E}_q\big[\,|e^{tX_c}/Z_t - 1|\,\big]$. Define

$$Y_t := \frac{e^{tX_c}}{Z_t}, \qquad R_t := Y_t - 1 - tX_c,$$

so that

$$\mathrm{TV}(p_t, q) = \frac{1}{2}\,\mathbb{E}_q\big[\,|tX_c + R_t|\,\big].$$

We will prove a uniform bound $\mathbb{E}_q[|R_t|] \le K(m,\rho)\, t^2$ with $K(m,\rho) < \infty$ depending only on (m,ρ). Then by the reverse triangle inequality,

$$\Big|\mathbb{E}_q[|tX_c + R_t|] - |t|\,\mathbb{E}_q[|X_c|]\Big| \le \mathbb{E}_q[|R_t|],$$

yielding

$$\Big|\,\mathrm{TV}(p_t,q) - \tfrac{|t|}{2}\,\mathbb{E}_q[|X_c|]\Big| \le \tfrac{1}{2}\,\mathbb{E}_q[|R_t|] \le \tfrac{1}{2}K(m,\rho)\, t^2.$$

Thus the two-sided inequality holds with the same constant $C(m,\rho) = \frac{1}{2}K(m,\rho)$, once we bound $K(m,\rho)$ explicitly.

From $R_t = e^{tX_c}/Z_t - 1 - tX_c$ we obtain

$$R_t = \frac{e^{tX_c} - 1 - tX_c}{Z_t} + \Big(\frac{1}{Z_t} - 1\Big)(1 + tX_c) = \frac{e^{tX_c} - 1 - tX_c}{Z_t} - \frac{Z_t - 1}{Z_t} - tX_c\,\frac{Z_t - 1}{Z_t}.$$

Since $e^y \ge 1+y$ for all y, $e^{tX_c} - 1 - tX_c \ge 0$ almost surely; and by Jensen, $Z_t = \mathbb{E}_q[e^{tX_c}] \ge e^{t\mathbb{E}_q[X_c]} = 1$. Therefore,

$$\mathbb{E}_q[|R_t|] \le \frac{1}{Z_t}\,\mathbb{E}_q[e^{tX_c} - 1 - tX_c] + \frac{Z_t - 1}{Z_t} + |t|\,\mathbb{E}_q[|X_c|]\cdot\frac{Z_t - 1}{Z_t}.$$

Let $\varepsilon := \mathrm{sgn}(t)$ and $A(t) := \mathbb{E}_q\big[e^{tX_c} - 1 - tX_c\big]$. Using Taylor's integral remainder,

$$A(t) = \int_0^{|t|} (|t| - s)\,\mathbb{E}_q\big[X_c^2 e^{s\varepsilon X_c}\big]\,ds.$$

Because $\mathbb{E}_q[X_c] = 0$, we also have $Z_t - 1 = \mathbb{E}_q[e^{tX_c} - 1] = A(t)$. Hence

$$\mathbb{E}_q[|R_t|] \le \frac{1}{Z_t}\big(2 + |t|\,\mathbb{E}_q[|X_c|]\big)A(t).$$

Bounding $Z_t \ge 1$ and $\int_0^{|t|}(|t| - s)ds = \frac{|t|^2}{2}$ gives

$$\mathbb{E}_q[|R_t|] \le \Big(1 + \tfrac{|t|}{2}\,\mathbb{E}_q[|X_c|]\Big)|t|^2 \sup_{0\le u\le |t|} \mathbb{E}_q\big[X_c^2 e^{u\varepsilon X_c}\big].$$

Next, uniformly for $u \in [\,0, \rho\,]$,

$$\mathbb{E}_q\big[X_c^2 e^{u\varepsilon X_c}\big] = e^{-u\varepsilon\mu} \sum_{v\in S} q(v)^{1+u\varepsilon}\big(\log q(v) - \mu\big)^2.$$

Using $e^{|u\mu|} \le e^{\rho|\mu|} \le m^\rho$, $(a-b)^2 \le 2a^2 + 2b^2$, and $\sum_v q(v)^{1+u\varepsilon} \le m^\rho$ (since for $\alpha \ge 1$, $\sum q^\alpha \le 1 \le m^\rho$, while for $\alpha \in (0,1)$, $\sum q^\alpha \le m^{1-\alpha} \le m^\rho$), we obtain

$$\mathbb{E}_q\big[X_c^2 e^{u\varepsilon X_c}\big] \le 2m^\rho\Big(\sum_v q(v)^{1+u\varepsilon}\log^2 q(v)\Big) + 2m^\rho\mu^2\sum_v q(v)^{1+u\varepsilon}.$$

Since $1 + u\varepsilon \ge 1 - \rho$ and, for $x \in (0,1]$, the map $\alpha \mapsto x^\alpha$ decreases in α, we have $x^{1+u\varepsilon} \le x^{1-\rho}$. Thus

$$\sum_v q(v)^{1+u\varepsilon}\log^2 q(v) \le |S| \cdot \sup_{x\in(0,1]} x^{1-\rho}(\log x)^2 \le m\,B_1(\rho),$$

where

$$B_1(\rho) := \sup_{x\in(0,1]} x^{1-\rho}(\log x)^2 = \sup_{y\ge 0} e^{-(1-\rho)y}y^2 = \frac{4}{e^2(1-\rho)^2}.$$

Using $|\mu| \le \log m$ yields the uniform bound

$$\sup_{0\le u\le\rho} \mathbb{E}_q\big[X_c^2 e^{u\varepsilon X_c}\big] \le 2m^\rho\big(mB_1(\rho) + (\log m)^2 m^\rho\big) =: M(m,\rho).$$

Finally, since $X \le 0$ almost surely, $\mathbb{E}_q[|X|] = -\mu \le \log m$, and hence $\mathbb{E}_q[|X_c|] \le \mathbb{E}_q[|X|] + |\mu| \le 2\log m$. Combining the pieces, for $|t| \le \rho$,

$$\mathbb{E}_q[|R_t|] \le \Big(1 + \tfrac{|t|}{2}\,\mathbb{E}_q[|X_c|]\Big)|t|^2 M(m,\rho) \le (1 + \rho\log m)\,M(m,\rho)\,t^2.$$

Therefore

$$\Big|\,\mathrm{TV}(p_t, q) - \frac{|t|}{2}\,\mathbb{E}_q[|X_c|]\Big| \le \frac{1}{2}\,\mathbb{E}_q[|R_t|] \le \frac{(1+\rho\log m)\,M(m,\rho)}{2}\,t^2.$$

Setting

$$C(m,\rho) := \frac{(1+\rho\log m)\,M(m,\rho)}{2} = (1+\rho\log m)\,m^\rho\big(mB_1(\rho) + (\log m)^2 m^\rho\big)$$

proves the claimed absolute deviation bound, hence the two-sided inequalities with the same constant $C(m,\rho)$, uniformly over steps and prefixes. In particular,

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

uniformly in n and $x_{1:n-1}$. □

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 $j \in \{0, \dots, K\}$,*

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

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

Proof. Fix an iteration index n and a realized history $h \equiv x_{1:n-1}$. For $\tau \in \{1, \dots, K\}$ let

$$P_p^{(\tau)}(x_{1:\tau} \mid h) \;=\; \prod_{i=1}^{\tau} p_{n+i-1}(x_i \mid h, x_{1:i-1}), \qquad P_q^{(\tau)}(x_{1:\tau} \mid h) \;=\; \prod_{i=1}^{\tau} q_{n+i-1}(x_i \mid h, x_{1:i-1})$$

denote the τ-step path laws of the proposal and target, respectively. Consider any unbiased (valid) K-block draft–verify scheme. By exactness, its (eventual) output law for the next K tokens given h is $P_q^{(K)}(\cdot \mid h)$. Therefore the scheme induces a coupling Γ_h of two random blocks $(X_{1:K}, Y_{1:K}) \in \mathcal{V}^K \times \mathcal{V}^K$ with marginals $P_p^{(K)}(\cdot \mid h)$ and $P_q^{(K)}(\cdot \mid h)$, where $X_{1:K}$ is the drafted block and $Y_{1:K}$ is the scheme's (eventual) output block.

Let the first-disagreement time be

$$\sigma \;:=\; \inf\{j \geq 1 : X_j \neq Y_j\} \;(\inf \emptyset := \infty), \qquad L^{(K)} \;:=\; \min\{K,\, \sigma - 1\}.$$

By construction of the induced coupling (accepted tokens are committed as the prefix of the output), the accepted-prefix length $A^{(K)}$ of the scheme satisfies $A^{(K)} \leq L^{(K)}$ almost surely, with equality $A^{(K)} = L^{(K)}$ for the longest-prefix (LP) verifier.

Hence, for every $j \in \{1, \dots, K\}$ and realized h,

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

The first inequality uses $\{A^{(K)} \geq j\} \subseteq \{X_{1:j} = Y_{1:j}\}$ under Γ_h; the second is the maximal-coupling bound applied to the distributions on $\mathcal{V}^j$.

Now average (1) over the random prefix $h \sim q$ (the target's prefix law): for any valid scheme,

$$\mathbb{P}(A^{(K)} \geq j) \;\leq\; \mathbb{E}_{h \sim q}\Big[\sum_{x_{1:j}} \min\{P_p^{(j)}(x_{1:j} \mid h), P_q^{(j)}(x_{1:j} \mid h)\}\Big] \;=:\; S_j. \tag{2}$$

Summing (2) over $j = 1, \dots, K$ and using $\mathbb{E}[A^{(K)}] = \sum_{j=1}^{K} \mathbb{P}(A^{(K)} \geq j)$ gives, for any valid scheme,

$$\mathbb{E}[A^{(K)}] \;\leq\; \sum_{j=1}^{K} S_j. \tag{3}$$

For the LP verifier, the block-verification optimality formula yields the exact value

$$\mathbb{E}[A_{\mathrm{LP}}^{(K)}] \;=\; \sum_{j=1}^{K} S_j. \tag{4}$$

Since each inequality in (2) is an individual upper bound with nonnegative slack and the sum of these slacks equals zero for LP by (3)–(4), every slack must be zero. Thus, for every $j \in \{1, \dots, K\}$,

$$\mathbb{P}_{\mathrm{LP}}\big(A^{(K)} \ge j\big) \;=\; S_j \;\ge\; \mathbb{P}_{\text{any valid}}\big(A^{(K)} \ge j\big).$$

For $j = 0$ the inequality is trivial. Hence the LP verifier first-order stochastically dominates every other unbiased K-block scheme, and in particular maximizes $\mathbb{E}[A^{(K)}]$. Equivalently, the LP-induced coupling attains the maximal agreement probabilities

$$\Gamma_h\big(X_{1:j} = Y_{1:j}\big) \;=\; 1 - \mathrm{TV}\left(P_p^{(j)}(\cdot \mid h), P_q^{(j)}(\cdot \mid h)\right) \quad \text{for all } j \in \{1, \dots, K\}. \qquad \square$$

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 tokenwise 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 $p_{n:n+K-1}$, let*

$$\pi_{n+i} := \mathrm{TV}\left(p_{n+i}(\cdot \mid X_{1:n+i-1}),\, q_{n+i}(\cdot \mid X_{1:n+i-1})\right), \qquad i = 0, \dots, K-1.$$

Then the full-acceptance probability satisfies the pathwise inequality

$$\prod_{i=0}^{K-1} \big(1 - \pi_{n+i}\big) \;\ge\; 1 - \sum_{i=0}^{K-1} \pi_{n+i}.$$

Averaging over prefixes yields the unconditional bound

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

Proof. Fix a block starting at position n of length K, and the longest-prefix verifier with proposal conditionals $(p_n, \dots, p_{n+K-1})$. For each $i \in \{0, \dots, K-1\}$ define the prefix-measurable total-variation gap

$$\pi_{n+i} \;=\; \mathrm{TV}\left(p_{n+i}(\cdot \mid X_{1:n+i-1}),\, q_{n+i}(\cdot \mid X_{1:n+i-1})\right) \in [0, 1].$$

By the maximal coupling lemma, at step $n + i$ and conditional on the event that the first i tokens in the block have been accepted (so both models condition on the same realized prefix $X_{1:n+i-1}$), there exists a coupling such that

$$\mathbb{P}\{\text{token } n+i \text{ is accepted} \mid X_{1:n+i-1},\, A^{(i)} = i\} \;=\; 1 - \pi_{n+i}.$$

Construct the joint coupling sequentially via the standard overlap–residual procedure, using fresh independent auxiliary randomness at each step. Then, by the chain rule of conditional probabilities,

$$\mathbb{P}\{A^{(K)} = K \mid X_{1:n+K-1}\} \;=\; \prod_{i=0}^{K-1} (1 - \pi_{n+i}).$$

We now use the elementary inequality: for any $a_0, \ldots, a_{K-1} \in [0, 1]$,

$$\prod_{i=0}^{K-1} (1 - a_i) \;\geq\; 1 - \sum_{i=0}^{K-1} a_i.$$

A short induction proves it: for $K = 1$ it is equality; if it holds for $K - 1$, then

$$\prod_{i=0}^{K-1} (1 - a_i) - \Big(1 - \sum_{i=0}^{K-1} a_i\Big) = \Big[\prod_{i=0}^{K-2} (1 - a_i) - \Big(1 - \sum_{i=0}^{K-2} a_i\Big)\Big] \;+\; a_{K-1}\Big(1 - \prod_{i=0}^{K-2} (1 - a_i)\Big) \;\geq\; 0.$$

Applying this pathwise with $a_i = \pi_{n+i}$ yields

$$\mathbb{P}\{A^{(K)} = K \mid X_{1:n+K-1}\} \;=\; \prod_{i=0}^{K-1} (1 - \pi_{n+i}) \;\geq\; 1 - \sum_{i=0}^{K-1} \pi_{n+i}.$$

Finally, taking expectations over the (target) prefix randomness and the internal coins gives

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

where $\mathbb{E}[\mathrm{TV}(p_{n+i}, q_{n+i})]$ abbreviates $\mathbb{E}_{X_{1:n+i-1} \sim q}\big[\,\mathrm{TV}(p_{n+i}(\cdot \mid X_{1:n+i-1}), q_{n+i}(\cdot \mid X_{1:n+i-1}))\big]$.

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 (i.e., the length of the initial run of matches within the 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, whenever $\gamma < \min_i \left(1 - \mathbb{E}[\mathrm{TV}(p_{n+i}, q_{n+i})]\right)$, every factor is strictly positive, and the right-hand side is strictly larger than the tokenwise worst-case guarantee (which corresponds to the vacuous replacement $\gamma \mapsto 1$, yielding zero).

Proof. Fix n and K. Assume

$$\forall x_{1:n}, x'_{1:n} : \mathrm{TV}\left(q_{n+1}(\cdot \mid x_{1:n}),\, q_{n+1}(\cdot \mid x'_{1:n})\right) \le \gamma\, \mathbf{1}\{x_n \ne x'_n\}, \qquad \gamma \in [0,1).$$

Work on the state space of prefixes: set $W_0 := X_{1:n-1}$ and, for $i \ge 0$, let $W_{i+1} := (W_i, X_{n+i})$ where $X_{n+i} \sim q_{n+i}(\cdot \mid W_i)$. Denote by $\mathsf{Q}_i(w,\cdot) := q_{n+i}(\cdot \mid w)$ the time-i transition on prefixes. By the assumption, if w, w' share the same last token then $\mathsf{Q}_i(w,\cdot) = \mathsf{Q}_i(w',\cdot)$; otherwise $\mathrm{TV}(\mathsf{Q}_i(w,\cdot), \mathsf{Q}_i(w',\cdot)) \le \gamma$. Hence the Dobrushin coefficient satisfies

$$\delta(\mathsf{Q}_i) := \sup_{w,w'} \mathrm{TV}\left(\mathsf{Q}_i(w,\cdot), \mathsf{Q}_i(w',\cdot)\right) \le \gamma.$$

For each in-block index $i \in \{0,\dots,K-1\}$ and prefix w, define the one-step maximal-agreement probability

$$U_i(w) \;:=\; 1 - \mathrm{TV}\left(p_{n+i}(\cdot \mid w),\; q_{n+i}(\cdot \mid w)\right) \in [0,1].$$

Couple the proposal and target processes by a Markovian maximal coupling across the block: at each step i, conditionally on the current coupled prefix states, draw the next pair of tokens by a maximal coupling of the corresponding conditionals.

Let B_i be the indicator that the i-th in-block tokens (position $n+i$) of the coupled proposal and target coincide, and set $T_i := \prod_{t=0}^{i} B_t \in \{0,1\}$ with the convention $T_{-1} \equiv 1$. The event that at least $j{+}1$ tokens are accepted is $\{T_j = 1\}$, so

$$\mathbb{P}\{A^{(K)} \ge j{+}1\} \;=\; \mathbb{E}[T_j], \qquad j = 0,1,\dots,K-1.$$

Crucially, we have the one-step recursion

$$\mathbb{E}[T_i] \;=\; \mathbb{E}[\,T_{i-1}\, U_i(W_i)\,], \qquad i \ge 0. \tag{5.1}$$

Indeed, conditionally on the entire past up to time i and on W_i, the Markovian maximal coupling ensures $\mathbb{P}\{B_i = 1 \mid T_{i-1} = 1, W_i\} = U_i(W_i)$, while $T_{i-1} = 0$ forces $T_i = 0$. Taking expectations yields (5.1).

We now lower bound $\mathbb{E}[T_i]$ using a one-step decoupling inequality for Markov chains.

Lemma 5.4 (one-step γ-decoupling)**.** *Let (W_i) be a (possibly time-inhomogeneous) Markov chain with transition Q_i satisfying $\delta(\mathsf{Q}_i) \le \gamma$. For any $i \ge 1$, any $\mathcal{F}_{i-1}$-measurable $Z \in [0,1]$ (where $\mathcal{F}_{i-1} := \sigma(W_0,\dots,W_{i-1})$), and any f : state $\to [0,1]$,*

$$\mathbb{E}[\,Z\, f(W_i)\,] \;\ge\; \mathbb{E}[Z], (\mathbb{E}[f(W_i)] - \gamma). \tag{5.2}$$

Proof of the lemma. Let $\mu_i := \mathcal{L}(W_i)$ be the unconditional law of W_i. By convexity of total variation, $\sup_w \mathrm{TV}(\mathsf{Q}_i(w,\cdot), \mu_i) \le \delta(\mathsf{Q}_i) \le \gamma$. Hence, for every $f \in [0,1]$ and all w,

$$\mathbb{E}[f(W_i) \mid W_{i-1} = w] \ge \mathbb{E}[f(W_i)] - \gamma.$$

Taking conditional expectation with respect to $\mathcal{F}_{i-1}$ and multiplying by any $\mathcal{F}_{i-1}$-measurable $Z \in [0,1]$ gives (5.2). □

Return to (**??**). Since T_{i-1} is $\mathcal{F}_{i-1}$-measurable, write $Z_i := \mathbb{E}[T_{i-1} \mid \mathcal{F}_{i-1}] = T_{i-1} \in [0,1]$. Then

$$\mathbb{E}[T_i] = \mathbb{E}[\,T_{i-1}\, U_i(W_i)\,] = \mathbb{E}[\,Z_i\, \mathbb{E}[U_i(W_i) \mid \mathcal{F}_{i-1}]] = \mathbb{E}[\,Z_i\, U_i(W_i)\,],$$

where the last equality uses Z_i being $\mathcal{F}_{i-1}$-measurable. Applying the lemma (5.2) with $f = U_i$ yields

$$\mathbb{E}[T_i] \;\ge\; \mathbb{E}[Z_i], (\mathbb{E}[U_i(W_i)] - \gamma) = \mathbb{E}[T_{i-1}], (\mathbb{E}[U_i(W_i)] - \gamma). \tag{5.3}$$

By induction on i (and noting $\mathbb{E}[T_0] = \mathbb{E}[U_0(W_0)] \ge \mathbb{E}[U_0(W_0)] - \gamma$), (5.3) gives

$$\mathbb{E}[T_j] \;\ge\; \prod_{i=0}^{j} (\mathbb{E}[U_i(W_i)] - \gamma), \qquad j = 0, 1, \dots, K-1.$$

Finally, by definition of U_i and taking expectations under the target prefix law and dynamics,

$$\mathbb{E}[U_i(W_i)] = 1 - \mathbb{E}[\mathrm{TV}(p_{n+i}, q_{n+i})].$$

Therefore

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

which is the claimed bound.

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].

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. Böettcher. 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 the 40th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, volume 202, pages 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 preprint arXiv:2203.16487, 2022. https://doi.org/10.48550/arXiv.2203.16487

[9] Moshe Shaked and J. George Shanthikumar. Stochastic Orders. Springer Series in Statistics. Springer, New York, 2007. https://doi.org/10.1007/978-0-387-34675-5

[10] Torgny Lindvall. Lectures on the Coupling Method. Dover Publications, Mineola, NY, 2002 (reprint of the 1992 Wiley edition). https://books.google.com/books?id=GUwyU1ypd1wC

[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. Preprint.

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