# WINNER-TAKE-ALL: WHERE EXPONENTIAL ENTROPY COLLAPSE COMES FROM IN REINFORCEMENT LEARNING FOR LLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) can improve accuracy while concentrating probability on fewer correct solutions. We study this process through a finite-mode, clean-verifier mean-field model of group relative policy optimization (GRPO), an algorithm used for RLVR. The correct-mode dynamics increase collision probability and decrease both Shannon and Rényi-2 entropy; a unique initial leader eventually captures all correct mass. We derive exact entropy–mass relations along a scalar trajectory and distinguish this mode-level concentration from token entropy through an entropy decomposition over solution families and token sequences. The distinction matters empirically: in a GRPO training run, correct-family diversity falls sharply early in training and only partially recovers, while token entropy subsequently rises above its initial level. Comparing unmodified GRPO with an anti-collision variant, family-count ratios remain between 4.4 and 5.9 over the tested training cutoffs, whereas token-entropy ratios cross unity. The scalar solution also characterizes when accuracy–entropy curves admit shifted-exponential or logistic approximations: each corresponds to a locally constant slope in a different transformed error coordinate. Fits to training curves across model families and sizes on math and coding tasks support both finite-window descriptions. These results motivate measuring correct-family collision alongside token entropy when assessing solution diversity in RLVR.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves language-model reasoning by rewarding solutions that pass an automated checker (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). These accuracy gains can accompany a loss of diversity: the policy assigns increasing probability to a small set of successful completions. Token-level policy entropy is a common diagnostic for this change, but it does not directly measure how many correct solution strategies remain. A model can vary its wording or response length while repeatedly using the same strategy. Understanding diversity loss therefore requires specifying the distribution whose entropy is being measured.

We analyze the concentration of correct solutions using the good–bad mean-field model of Rad et al. (2026b). For a fixed prompt, the policy places mass on finitely many correct and incorrect modes. The conditional probabilities of correct modes follow a collision flow: an initially more probable mode grows relative to a less probable one. Building on the known concentration and limiting-support results, we characterize the associated Shannon and Rényi-2 entropies. The scalar parameterization then gives exact relations between error mass, concentration, and entropy.

The distinction between mode entropy and token entropy is essential to interpreting this result. Family-level entropy measures variation between solutions; sequence entropy also includes variation within each family. Average token entropy further depends on sequence length and on how prefixes are weighted. We make these relations explicit and examine the corresponding training measurements. Figure 1 illustrates the central observation: token entropy can recover substantially while the sampled correct solutions remain concentrated.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

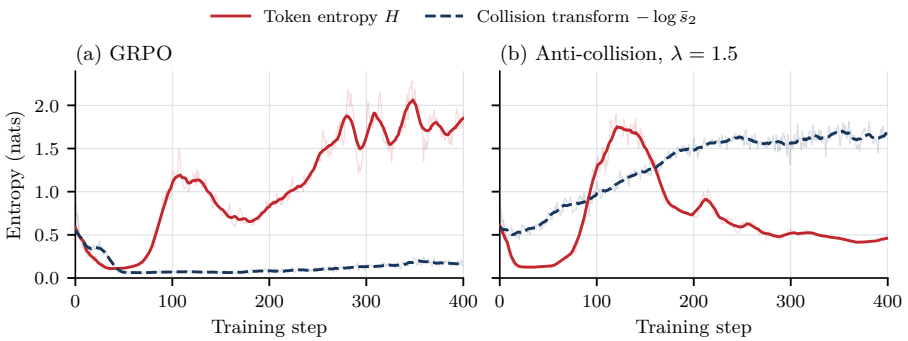


Figure 1: Token entropy and correct-family concentration in Qwen3-14B-Base. The collision-derived summary $-\log \bar{s}_2$ and logged token entropy are shown in nats. Under unmodified GRPO, the initial concentration is followed by partial recovery; token entropy later exceeds its initial level. The anti-collision run has a different family-diversity trajectory despite a nonmonotone token-entropy trace. Faint lines show logged values; thick lines show centered 15-observation means. Batch statistics and the intervention gain $\lambda$ are defined in Section 5.

In GRPO runs (Rad et al., 2026a), an anti-collision variant that favors underrepresented correct families produces roughly five times as many observed families as unmodified GRPO on Qwen3-14B-Base over a range of training cutoffs. The token-entropy comparison changes direction within that range, motivating the joint monitoring of both quantities.

**Relation to prior work.** Cui et al. (2025) describe a shifted-exponential relation between aggregate entropy and accuracy. Other work studies high-entropy decision tokens (Wang et al., 2025), entropy-dependent training objectives (Cheng et al., 2025; Vanlioglu, 2025; Deng et al., 2025), and confidence-based optimization (Agarwal et al., 2025; Prabhudesai et al., 2025). Our analysis connects these entropy diagnostics to concentration within the correct solution distribution. We build on the collision dynamics and scalar parameterization of Rad et al. (2026b) to characterize conditional entropy, derive entropy–mass relations, and bound the error of finite-window accuracy–entropy approximations. The family-level diagnostic and anti-collision method of Rad et al. (2026a) provide the measurements and intervention used in our empirical comparison.

**Contributions.**

- **We characterize Shannon and Rényi-2 entropy along the collision flow.** The analysis separates concentration within each block from the loss of incorrect mass and accounts for ties between dominant modes (Section 3).
- **We relate family, sequence, and token entropy.** We show that token entropy and observed correct-family diversity can rank two GRPO variants differently (Sections 4–5).
- **We explain when and why exponential accuracy–entropy relations arise.** From the coupled entropy–mass trajectory, we derive finite-window conditions and error bounds for exponential and logistic approximations and identify a non-exponential terminal entropy law (Section 7).

## 2 Solution families and the GRPO flow

For a fixed prompt, a completion receives a verifier reward after generation. Treating the complete response as one action gives a one-step reward problem, or bandit. This view separates improving accuracy from concentrating probability on fewer successful solution strategies.

We represent the completion distribution by $K$ correct modes and $M$ incorrect modes. A mode can represent one completion or a family of completions that share a solution strategy. We assume a fixed set of active modes, a noiseless verifier with reward one for correctness and zero otherwise, and an independently adjustable score, or logit, for each mode. We study unregularized, infinitesimal

updates using population averages, excluding clipping and finite-group noise. For semantic families, independent logits are a modeling assumption: shared neural parameters can produce different family-level updates.

Let $p \in (0,1)$ be the probability of an incorrect completion. Conditioned on correctness, $y_j$ is the probability of correct mode $j$; conditioned on an error, $z_m$ is the probability of incorrect mode $m$. The full policy is therefore

$$\pi = \big((1-p)y_1, \dots, (1-p)y_K,\ pz_1, \dots, pz_M\big), \qquad \sum_j y_j = \sum_m z_m = 1. \tag{2.1}$$

Thus $1-p$ measures accuracy, while $y$ describes how probability is shared among correct solutions.

The collision probability $s_2 = \sum_j y_j^2$ is the chance that two independently sampled correct completions belong to the same mode; $t_2 = \sum_m z_m^2$ is its incorrect-block counterpart. Larger collision means greater concentration. We distinguish Shannon entropy $H_y = -\sum_j y_j \log y_j$ from Rényi-2 entropy $H_2(y) = -\log s_2$. Uniform use of $K$ correct modes gives $s_2 = 1/K$, motivating the effective number $K_{\text{eff}} = 1/s_2 \le K$. The number of clusters observed in a finite rollout group is a separate, sample-dependent quantity.

A rollout group contains completions sampled for the same prompt. GRPO subtracts the group mean reward and divides by the group standard deviation to obtain each completion's advantage. We replace these random group statistics by their population values. Binary rewards then have mean $1-p$ and standard deviation $\varsigma(p) = \sqrt{p(1-p)}$, giving conditional advantages

$$\mathcal{A}_{\text{good}} = \frac{p}{\varsigma(p)}, \qquad \mathcal{A}_{\text{bad}} = -\frac{1-p}{\varsigma(p)}. \tag{2.2}$$

Equal advantages need not give equal logit increments. The mode logits define a softmax policy, $\pi_i = e^{\theta_i}/\sum_\ell e^{\theta_\ell}$, whose Jacobian is $\mathfrak{J}(\pi) = \mathrm{Diag}(\pi) - \pi\pi^\top$. Let $\mathcal{A}$ collect the mode advantages above. Holding them fixed when computing each policy gradient gives $\dot\theta = \eta\mathfrak{J}(\pi)\mathcal{A}$, where $\eta > 0$ is the learning-rate scale. Since $\pi^\top\mathcal{A} = 0$, this reduces to $\dot\theta_i = \eta\pi_i\mathcal{A}_i$: within the correct block, a more probable mode receives a larger expected logit increment. Applying the chain rule gives $\dot\pi = \mathfrak{J}(\pi)\dot\theta = \eta\mathfrak{J}(\pi)^2\mathcal{A}$. As derived by Rad et al. (2026b), the resulting block dynamics are

$$\dot y = \kappa(p)\big(y \odot y - s_2 y\big), \qquad \dot z = -\kappa(p)\big(z \odot z - t_2 z\big), \tag{2.3}$$

$$\dot p = -\eta[p(1-p)]^{3/2}(s_2 + t_2), \qquad \kappa(p) = \eta\sqrt{p(1-p)}. \tag{2.4}$$

Here dots denote derivatives with respect to training time $t$, and $\odot$ denotes elementwise multiplication. In particular, $\dot y_j/y_j = \kappa(p)(y_j - s_2)$: a correct mode gains conditional probability when its share exceeds $s_2$, while the total incorrect mass $p$ decreases.

Introducing internal time $\tau(t) = \int_0^t \kappa(p(s))\ \mathrm{d}s$ removes the common speed factor without changing the probability trajectory. For example, with two correct modes, $y = (u, 1-u)$, the first coordinate satisfies $\mathrm{d}u/\mathrm{d}\tau = u(1-u)(2u-1)$. A 60:40 split becomes more unequal, whereas an exact 50:50 tie remains fixed. Equal rewards therefore do not imply equal growth of correct-mode probabilities.

## 3 Entropy collapse in the collision flow

With dots now denoting derivatives with respect to $\tau$, the system becomes

$$\dot y = y \odot (y - s_2\mathbf{1}), \qquad \dot z = z \odot (t_2\mathbf{1} - z), \qquad \dot p = -p(1-p)(s_2 + t_2). \tag{3.1}$$

Building on the correct-block concentration result of Rad et al. (2026b), the following theorem characterizes both conditional entropies and separates within-block concentration from the loss of total incorrect mass. Here and below, entropy uses natural logarithms.

*Theorem* 3.1 (Entropy and limiting support). Let $K, M$ be the finite positive sizes of the initial active supports, let $p(0) \in (0,1)$, and let $y(0) = q$ and $z(0) = r$ be positive on their active supports. Along (3.1):

(a) Correct-mode collision increases and incorrect-mode collision decreases:

$$\dot{s}_2 = 2\sum_j y_j(y_j - s_2)^2 \geq 0, \qquad \dot{t}_2 = -2\sum_m z_m(z_m - t_2)^2 \leq 0. \tag{3.2}$$

Consequently, $H_2(y)$ is nonincreasing and $H_2(z)$ is nondecreasing.

(b) The conditional Shannon entropies satisfy

$$\dot{H}_y = -\operatorname{Cov}_{j\sim y}(y_j, \log y_j) \leq 0, \qquad \dot{H}_z = \operatorname{Cov}_{m\sim z}(z_m, \log z_m) \geq 0. \tag{3.3}$$

(c) Let $\mathcal{M}_\star = \{j : q_j = \max_i q_i\}$ have size $m_\star$. Then $y(\tau) \to m_\star^{-1}\mathbf{1}_{\mathcal{M}_\star}$, $z(\tau) \to M^{-1}\mathbf{1}$, and $p(\tau) \to 0$ as $\tau \to \infty$. Both correct-block entropies converge to $\log m_\star$. The full mode entropy

$$H_{\text{mode}} = h_{\text{bin}}(p) + (1-p)H_y + pH_z$$

also converges to $\log m_\star$, where $h_{\text{bin}}(p) = -p\log p - (1-p)\log(1-p)$. A unique initial correct-mode leader therefore yields a correct, single-mode limiting policy.

**Proof sketch.** Differentiating $s_2$ and $t_2$ gives (3.2). Differentiating Shannon entropy gives (3.3); the covariance is nonnegative because $x$ and $\log x$ are increasing on the active support. The log-ratio identities

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\log\frac{y_i}{y_j} = y_i - y_j, \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}\log\frac{z_m}{z_\ell} = -(z_m - z_\ell) \tag{3.4}$$

preserve coordinate order. The scalar solution in Section 7 identifies the limiting supports. Finally, $\mathrm{d}\operatorname{logit}(p)/\mathrm{d}\tau = -(s_2 + t_2)$ and $s_2 + t_2 \geq 1/K + 1/M$ imply $p \to 0$. The full proof is in Appendix C.4.

The theorem describes two different processes: correct modes concentrate, while incorrect modes spread conditionally but lose total probability. Full mode entropy need not decrease at every intermediate time. Its limit follows because the contribution $pH_z$ vanishes and the surviving correct block concentrates. These are structural properties of the specified mean-field model; their empirical counterparts must be assessed separately in neural training.

## 4 From solution entropy to token entropy

Token entropy and correct-family entropy summarize different sources of variation. Let $X$ be a complete token sequence, including termination, and let $F = f(X)$ assign each correct sequence to a solution family. For a fixed partition,

$$H(X \mid \text{correct}) = H_y + H(X \mid F, \text{correct}). \tag{4.1}$$

The second term captures variation within a family, such as different wording or intermediate steps. Thus even the Shannon entropy of complete correct sequences is generally larger than $H_y$. The collision entropy $H_2(y)$ is a different functional again, although the theorem establishes the same monotonicity for both family entropies.

A prefix tree gives the corresponding token-level accounting. For a finite sequence distribution, let $m_h$ be the probability of reaching prefix $h$, let $\pi(a \mid h)$ be the next-token distribution, and let $H_h = -\sum_a \pi(a \mid h)\log\pi(a \mid h)$. The chain rule gives

$$H(X) = \sum_{h\in\mathcal{T}_{\text{int}}} m_h H_h, \qquad \bar{H}_{\text{tok}} = \frac{H(X)}{\mathbb{E}[|X|]}. \tag{4.2}$$

The second identity uses token-weighted averaging. Other logging conventions, including averaging a length-normalized value per sequence, need not give this ratio. Logged training entropy also includes incorrect sequences, whereas $H_y$ conditions on correctness.

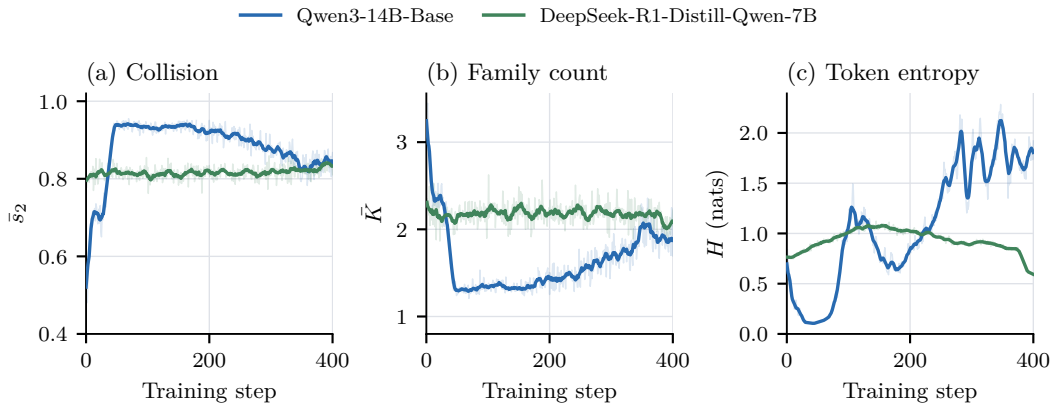


Figure 2: Correct-family concentration in two unmodified GRPO runs. Qwen3-14B-Base shows a marked early rise in collision and a decline in family count, followed by partial recovery. DeepSeek-R1-Distill-Qwen-7B starts from a more concentrated policy. Curves show centered 9-observation means over faint logged values. The runs differ in size and pretraining, so this comparison does not isolate either factor.

When each mode is a single sequence, prefix probabilities follow directly from $y$. If $\mathcal{L}(h)$ is the set of correct leaves below $h$, then

$$m_h = \sum_{j \in \mathcal{L}(h)} y_j, \qquad \pi(a \mid h) = \frac{m_{ha}}{m_h} \quad (m_h > 0). \tag{4.3}$$

These are probabilities conditional on correctness. With a unique winning leaf, the next-token distribution becomes deterministic on its path. Prefixes outside that path have vanishing occupancy, even when their local entropy remains positive (Appendix F). This single-sequence interpretation is an additional assumption; concentration on a semantic family alone does not imply deterministic tokens within that family.

Equations (4.1)–(4.2) show why token entropy alone cannot determine family diversity. Within-family variation, correctness, and length all contribute. Even when sequence entropy decreases, average token entropy can rise if expected length decreases proportionally faster. Identifying the cause of an observed difference requires further measurements.

## 5 Measuring Correct-Solution Diversity

We examine training panels from DeepSeek-R1-Distill-Qwen-7B and Qwen3-14B-Base (Rad et al., 2026a). The unmodified runs use GRPO with $G = 16$ rollouts per prompt and no KL or explicit diversity term. Correct rollouts are embedded and clustered within each prompt using cosine-distance DBSCAN (Ester et al., 1996), with noise points treated as singleton families. For cluster counts $n_j$ among $n$ correct rollouts, we write $\hat{y}_j = n_j/n$ and measure $\hat{s}_2 = \sum_j \hat{y}_j^2$ and the occupied-cluster count $\hat{K}$. Clusters approximate solution families without establishing semantic equivalence. Bars denote batch means over prompts with at least two correct rollouts; batches vary during training. The plotted $-\log \bar{s}_2$ transforms mean collision, rather than averaging per-prompt Rényi entropy. We report descriptive comparisons through optimizer step 400; Appendix A gives the protocol.

**Concentration under unmodified GRPO.** Figure 2 shows early concentration on Qwen3-14B-Base: smoothed correct-family collision rises from approximately $0.52$ to $0.94$, while family count falls from $3.25$ to $1.29$ within 68 optimizer steps. Diversity subsequently recovers in part. The initial phase agrees with the model's concentration mechanism, while the later recovery shows that its monotone trajectory does not describe the entire neural training run. The distilled model begins closer to a concentrated distribution and exhibits smaller changes.

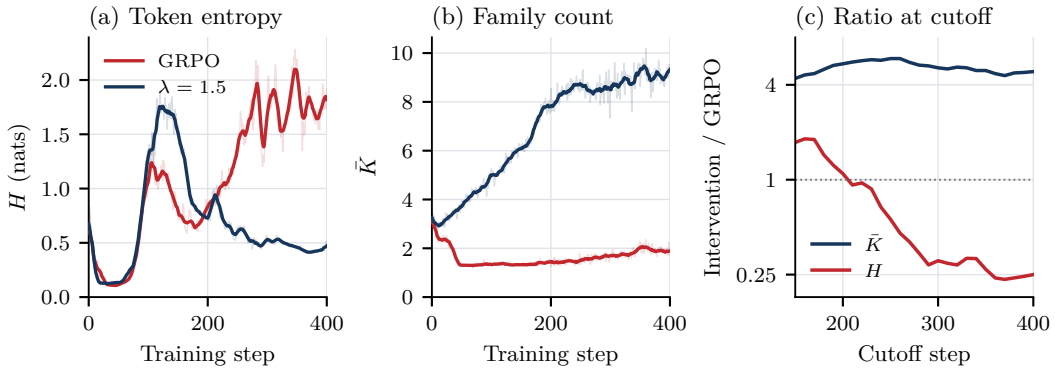


Figure 3: Family-count and token-entropy comparisons between unmodified GRPO and an anti-collision run on Qwen3-14B-Base. Training traces show centered 11-observation means over faint logged values. Ratios use the median of the last 25 observations at each cutoff from step 150 to 400. Family-count ratios remain on the same side of unity, whereas the token-entropy comparison changes direction.

**Anti-collision intervention.** G$^2$RPO (Rad et al., 2026a) adds a family-dependent bonus to the GRPO advantage. Its coefficient $g \geq 0$ is scheduled from group statistics using the configured gain $\lambda$ shown in the figures; $\lambda = 0$ recovers GRPO. For a correct rollout $i$ in family $j$, the unclipped form is

$$\widehat{\mathcal{A}}_i^{\mathrm{AC}} = \widehat{\mathcal{A}}_i^{\mathrm{GRPO}} + B_i, \qquad B_i = g\left(\frac{1}{\hat{K}\hat{y}_j} - 1\right). \tag{5.1}$$

For $g > 0$, families below the uniform share $1/\hat{K}$ receive positive adjustments, and those above it receive negative adjustments. The bonus is broadcast across response tokens. Appendix A.1 describes its scaling and the optional adjustment for incorrect completions.

**Token entropy does not order family diversity consistently.** We compare the base-model run with its anti-collision counterpart, using the same backbone and rollout group size. Across cutoffs from step 150 to 400, the intervention-to-unmodified ratio of correct-family counts lies between 4.4 and 5.9. Over the same cutoffs, the corresponding token-entropy ratio crosses unity near step 200 (Figure 3). The ordering by token entropy therefore depends on the evaluation point, whereas the family-count ordering is stable over these cutoffs. This comparison supports measuring the two quantities separately; it does not identify the cause of the token-entropy changes.

## 6 Implications for Diversity Control

The reciprocal bonus opposes the within-correct collision field in the population model. Sufficient effective gain reverses this field and hence the sign of $\dot{s}_2$ (Appendix A.1). In contrast, multiplying the entire flow by a positive constant preserves its trajectories and limiting support. General KL penalties and entropy bonuses need not act as such rescalings.

Figure 4 shows lower collision and larger family counts under the tested positive gains on both backbones. The observed diversity gains do not imply that accuracy increases with $\lambda$; Appendix A.3 reports the available validation traces. A separate 17-run panel covers entropy bonuses, KL penalties, covariance-based variants, and diversity bonuses (Appendix A.2). Its correlations describe the measurements under those settings rather than establish an unchanged collision mechanism.

For practical monitoring, we recommend reporting correct-family collision, observed family count, and the number of correct samples alongside token entropy. These measurements require no additional rollouts, although embedding and clustering add computation. Reporting sample counts is necessary because both collision estimates and observed family counts depend on the number of correct completions.

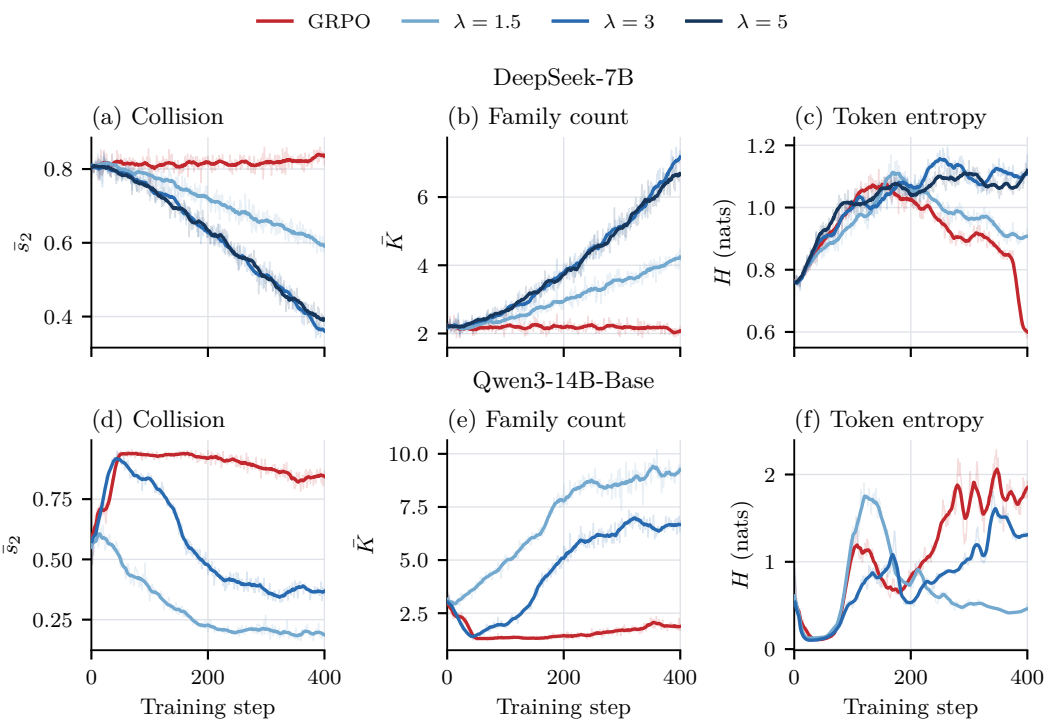


Figure 4: Correct-family statistics and token entropy for anti-collision configurations with gain $\lambda$ (Rad et al., 2026a). The tested positive gains generally yield lower collision and more observed families than unmodified GRPO ($\lambda = 0$). Token entropy does not consistently preserve this ordering. Faint lines show logged values; thick lines show centered 15-observation means.

## 7 Finite-window accuracy–entropy relations

The scalar trajectory gives an exact parametric relation between entropy and accuracy. We first derive this relation, then bound the error of its exponential and logistic approximations. Throughout the derivation, $H = H_y$ denotes correct-mode Shannon entropy and $R = 1 - p$ denotes accuracy.

### 7.1 Exact entropy–mass representation

Let $q = y(0)$, $r = z(0)$, and $p_0 = p(0) \in (0, 1)$, with $q$ and $r$ positive on their active supports. Applying the reciprocal parameterization of Rad et al. (2026b) to both blocks gives

$$y_j(I) = \frac{q_j/(1 - Iq_j)}{A(I)}, \quad z_m(U) = \frac{r_m/(1 + Ur_m)}{A_-(U)}, \qquad \begin{aligned} A(I) &= \sum_j \frac{q_j}{1 - Iq_j}, \\ A_-(U) &= \sum_m \frac{r_m}{1 + Ur_m}. \end{aligned} \tag{7.1}$$

The clocks satisfy $\mathrm{d}I/\,\mathrm{d}\tau = 1/A(I)$ and $\mathrm{d}U/\,\mathrm{d}\tau = 1/A_-(U)$, with $I(0) = U(0) = 0$. Matching their internal times gives

$$-\sum_j \log(1 - Iq_j) = \sum_m \log(1 + U(I)r_m), \qquad 0 \le I < 1/q_{\max}, \tag{7.2}$$

where $q_{\max} = \max_j q_j$. The right-hand side increases continuously from zero to infinity, so each admissible $I$ determines a unique $U(I) \ge 0$.

**Proposition 7.1** (Exact entropy–mass curve)**.** *With $U(I)$ defined by* (7.2)*, the trajectory satisfies*

$$p(I) = \frac{p_0 A_-(U(I))}{(1 - p_0)A(I) + p_0 A_-(U(I))}, \tag{7.3}$$

$$H(I) = \log A(I) - \sum_j y_j(I) \log \frac{q_j}{1 - I q_j}. \tag{7.4}$$

*Thus $(H(I), 1 - p(I))$ parameterizes the accuracy–entropy curve.*

**Derivation.** The normalizers obey $\mathrm{d}\log A/\,\mathrm{d}\tau = s_2$ and $\mathrm{d}\log A_-/\,\mathrm{d}\tau = -t_2$. Integrating $\mathrm{d}\,\mathrm{logit}\, p/\,\mathrm{d}\tau = -(s_2 + t_2)$, with $A(0) = A_-(0) = 1$, yields $\mathrm{logit}\, p = \mathrm{logit}\, p_0 - \log A + \log A_-$ and hence (7.3). Substituting $\log y_j = \log q_j - \log(1 - I q_j) - \log A$ into Shannon entropy gives (7.4). Appendix C supplies the full derivation.

The same calculation gives the local slope without inverting $H(I)$. Write $\mathcal{S} = s_2$, $\mathcal{T} = t_2$, and $\mathcal{C} = \mathrm{Cov}_{j\sim y}(y_j, \log y_j)$. The scalar derivatives are

$$\frac{\mathrm{d}H}{\mathrm{d}I} = -A\mathcal{C}, \qquad \frac{\mathrm{d}p}{\mathrm{d}I} = -Ap(1-p)(\mathcal{S} + \mathcal{T}). \tag{7.5}$$

Since $\mathrm{d}p/\,\mathrm{d}I < 0$, their ratio gives

$$\frac{\mathrm{d}H}{\mathrm{d}p} = \frac{\mathcal{C}}{p(1-p)(\mathcal{S} + \mathcal{T})}. \tag{7.6}$$

The curve therefore depends on both within-correct concentration and the changing correct/incorrect mass split. The scalar speed factor cancels, but the initial distributions remain in the relation through $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{C}$.

### 7.2 Exponential and logistic approximations

On a nonuniform correct-block trajectory, $\mathcal{C} > 0$ at finite internal time. The transformed error coordinates then satisfy

$$\frac{\mathrm{d}\,\mathrm{logit}\, p}{\mathrm{d}H} = \frac{\mathcal{S} + \mathcal{T}}{\mathcal{C}} =: \Lambda(H), \qquad \frac{\mathrm{d}\log p}{\mathrm{d}H} = (1-p)\Lambda(H) =: \mathcal{B}(H). \tag{7.7}$$

We refer to these slopes as the logit and residual-exponential hazards; the term denotes a slope diagnostic, without a survival-model assumption. For a reference entropy $H_0$, write $\Delta H = H - H_0$. Choose constant slopes $B_0$ and $L_0$, for example $\mathcal{B}(H_0)$ and $\Lambda(H_0)$, and define

$$p_{\exp}(H) = p(H_0) e^{B_0 \Delta H}, \tag{7.8}$$

$$R_{\log}(H) = \left[1 + e^{\mathrm{logit}\, p(H_0) + L_0 \Delta H}\right]^{-1}. \tag{7.9}$$

Since $R = 1 - p$, the first gives a free-slope shifted exponential for accuracy. The unit-slope form $1 - ae^H$ additionally requires $B_0 \approx 1$ in the chosen entropy units.

**Corollary 7.2** (Finite-window approximation error)**.** *Let $J$ be a compact interval within the entropy range attained at finite internal times on a nonuniform trajectory, with $H_0 \in J$. Define $\delta_B = \sup_{u\in J} |\mathcal{B}(u) - B_0|$ and $\delta_L = \sup_{u\in J} |\Lambda(u) - L_0|$. For every $H \in J$,*

$$\left|\log \frac{p(H)}{p_{\exp}(H)}\right| \le \delta_B |\Delta H|, \qquad |R(H) - R_{\log}(H)| \le \frac{\delta_L}{4} |\Delta H|. \tag{7.10}$$

**Derivation.** Integrating $\mathcal{B} - B_0$ bounds the log-relative error in $p$. Integrating $\Lambda - L_0$ bounds the logit error; the logistic function has derivative at most $1/4$, giving the accuracy bound. Thus slope variation and entropy-window width jointly control the approximation error. Appendix G gives the corresponding exponential and logistic envelopes.

The terminal limit rules out these constant-slope formulas as exact global relations. With at least two active correct modes, a unique initial leader, and $M$ active incorrect modes, the terminal asymptotic is

$$H(p) \sim C p^{\alpha} |\log p|, \qquad \alpha = \frac{M}{M+1}, \quad C > 0, \tag{7.11}$$

as $p \downarrow 0$ (Appendix D). This algebraic-logarithmic tail is distinct from a global exponential relation. Here $M$ is the active incorrect support size; an initial inverse collision is not generally an interchangeable quantity.

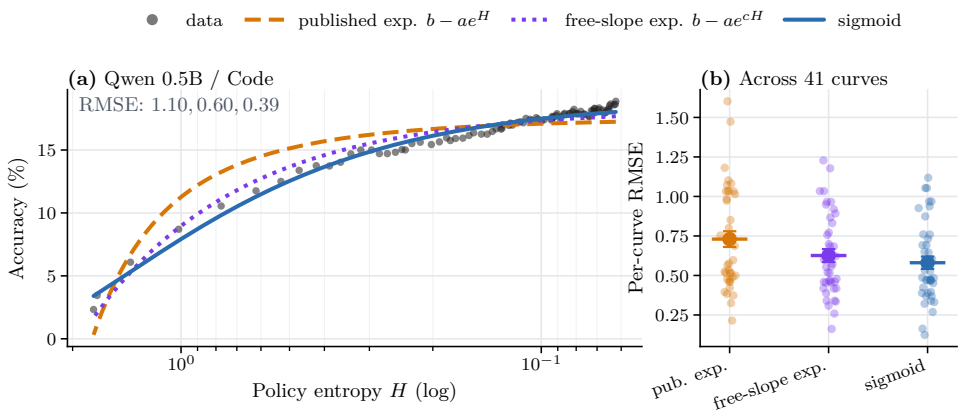


Figure 5: Accuracy–entropy fits with a free-slope exponential baseline. Left: representative fits. Right: per-curve RMSE, measured in validation accuracy percentage points. The sigmoid has lower mean residual error, while the free-slope exponential is preferred more often after the parameter penalty (Appendix B.1).

### 7.3 Comparison on published training curves

We compare the resulting curve families on 41 entropy–accuracy traces from Cui et al. (2025). These traces report aggregate logged entropy rather than $H_y$; here $H$ denotes that logged quantity. The comparison therefore assesses curve shape without identifying the underlying mode dynamics. We fit the published shifted exponential $b - ae^{H}$, a free-slope exponential $b - ae^{cH}$, and a four-parameter saturating sigmoid independently to each trace. Free vertical offsets and sigmoid saturation levels extend the theoretically calibrated forms. Appendix B describes the fits, complexity penalties, and held-out windows.

Mean RMSE is $0.730$, $0.622$, and $0.581$ percentage points for the published exponential, free-slope exponential, and sigmoid, respectively. The sigmoid has the lower descriptive AICc score than the published form on $33/41$ curves, but the free-slope exponential has the lowest score among all three forms on $27/41$. Fitting the first 60% of each trajectory and evaluating on the remaining 40% gives mean RMSEs of $0.882$, $0.876$, and $0.855$. The modest differences and mixed preferences support using finite-window approximations rather than selecting one universal curve. The exact hazard identities explain the conditions for each approximation; aggregate fit quality alone cannot establish those conditions for an individual prompt.

## 8 Discussion

The theory establishes correct-mode concentration under a fixed-support, clean-verifier mean-field model. Token sharing, finite rollout groups, clipping, regularization, and verifier error can alter these dynamics. Empirically, embedding clusters approximate solution families, and the available runs do not separate model size, pretraining, or seed variation. The early concentration and entropy-ordering reversal provide evidence for monitoring family diversity separately from token entropy; they do not establish monotone collapse throughout neural training.

The main implication is that entropy measurements must be interpreted at the level of the distribution they summarize. The collision flow specifies how correct-mode diversity is lost, the scalar solution characterizes its accuracy–entropy geometry, and the sequence decomposition explains why logged token entropy need not follow the same trajectory. Correct-family collision complements token entropy by measuring concentration among the solutions that receive reward.

## REPRODUCIBILITY STATEMENT

Appendices C–D provide the scalar derivation, theorem proofs, and terminal asymptotics. Appendix A describes the rollout measurements, aggregation, and analysis windows. Appendix B specifies the curve-fitting models and comparisons on the traces of Cui et al. (2025).

## STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

Large language models assisted with manuscript revision, data analysis and figure code, and checking and revising mathematical arguments. The authors take responsibility for the final text, claims, figures, and code.

## REFERENCES

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in LLM reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. *arXiv preprint arXiv:2505.22617*, 2025.

Jia Deng, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, and Ji-Rong Wen. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning. *arXiv preprint arXiv:2508.02260*, 2025.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231. AAAI Press, 1996.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*, 2025.

Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. $G^2$rpo: Geometric grpo: Escaping llms' reasoning ruts to break the accuracy–entropy trade-off. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *PMLR*, 2026a.

Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. Rate or fate? rlver: Reinforcement learning with verifiable noisy rewards. *arXiv preprint arXiv:2601.04411*, 2026b.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Abdullah Vanlioglu. Entropy-guided sequence weighting for efficient exploration in rl-based llm fine-tuning. *arXiv preprint arXiv:2503.22456*, 2025.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

# A EMPIRICAL PROTOCOL AND ADDITIONAL COMPARISONS

This appendix specifies the rollout statistics used in Sections 5 and 6, followed by additional comparisons across training settings.

## A.1 PROTOCOL

**Training settings and panels.** The two unmodified GRPO runs use DAPO-17K, learning rate $10^{-6}$, $G = 16$ rollouts per prompt, and no diversity bonus, KL penalty, or entropy bonus (Rad et al., 2026a). The training batch contains 256 prompts, with some base-model segments using 128. The saved histories contain 550 and 626 observations for DeepSeek-R1-Distill-Qwen-7B and Qwen3-14B-Base, respectively. We retain observations through step 400. The regularization panel in Appendix A.2 contains 17 DeepSeek-7B runs under additional training settings. The comparisons do not constitute independent seed replicates of a common configuration.

**Family definition and aggregation.** In the two unmodified runs, correct rollouts are embedded with all-MiniLM-L6-v2 from Sentence Transformers (Reimers & Gurevych, 2019), $L_2$-normalized, and clustered separately within each prompt using DBSCAN with cosine distance and `min_samples=2`. The source protocol specifies a radius of $\varepsilon = 0.05$ (Rad et al., 2026a). Noise points form singleton families. Prompts with fewer than two correct rollouts are excluded from the family panel. Collision and family count are first computed per retained prompt and then averaged over the batch. In the unmodified runs these statistics are observational; the anti-collision variants also use family statistics in their training bonus.

For a prompt with $n$ correct rollouts and cluster counts $n_j$, the reported statistics are

$$\hat{s}_2 = \sum_j (n_j/n)^2, \qquad \hat{K} = \#\{j : n_j > 0\}.$$

We write $\bar{s}_2$ for the batch mean of $\hat{s}_2$. The plotted entropy summary $-\log \bar{s}_2$ is the negative logarithm of that mean, not the mean of per-prompt Rényi entropies. Likewise, $1/\bar{s}_2$ is not generally the mean effective family count. Token entropy is the separately logged `actor/entropy` statistic. The varying training batches do not track a fixed prompt's latent distribution through time.

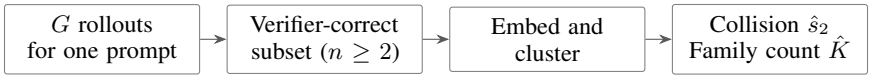


Figure 6: Construction of the correct-family panel. Statistics are computed within each retained prompt and then averaged over the training batch. Clustering is an operational definition of a family.

**Sampling and selection.** For fixed family assignments and independent samples, an unbiased collision estimator is

$$\tilde{s}_2 = \frac{\sum_j n_j(n_j - 1)}{n(n-1)} = \frac{n\hat{s}_2 - 1}{n-1}, \qquad n \geq 2.$$

The available batch summaries do not permit this correction prompt by prompt. A correction using $n \approx G(1 - \bar{p})$ is only an aggregate approximation. Data-dependent clustering adds a further source of estimation error. Observed family counts also depend on the number of correct samples, so they should be interpreted together with accuracy and panel coverage rather than as population support sizes.

**Windows and summaries.** Training traces show the raw logged values and centered rolling means. The concentration figure uses 9 observations, the cutoff-comparison figure uses 11, and the remaining main trajectory figures use 15. These widths count observations, which need not be equally spaced in optimizer steps. Cutoff ratios use medians over the last 25 available observations up to each cutoff. Correlations are descriptive Spearman coefficients on the unsmoothed values. We do not treat neighboring training steps as independent replicates or report per-step significance tests.

**Clustering sensitivity.** A solution family here is an embedding cluster, so its count depends on both the representation and the distance threshold. In training ablations on DeepSeek-R1-Distill-Qwen-7B, Rad et al. (2026a) halve the DBSCAN radius from $0.05$ to $0.025$ and replace MiniLM with bge-small-en-v1.5 in separate comparisons. Both changes preserve the increasing family-count trend and reach the same reported evaluation accuracy at step 130, while absolute counts differ. A separate 100-completion diagnostic yields 5–7 clusters for radii between $0.04$ and $0.11$. These checks support interpreting diversity trends under a fixed protocol; absolute counts remain specific to the embedding and clustering resolution.

**Incorrect-mode measurements.** The incorrect-block panel uses inconsistent placeholder values when a prompt has no incorrect rollouts. Depending on the scorer, such prompts contribute either zero or one to the logged collision statistic. We therefore do not use these aggregates to test the theorem's predictions for $t_2$ or $H_z$. Correct-family measurements are computed on the retained prompts with at least two correct rollouts.

**Anti-collision bonus and gain.** The core G$^2$RPO construction adds a reciprocal-frequency bonus to the normalized GRPO advantage, broadcast across the response tokens (Rad et al., 2026a). To relate this construction to the collision field, first omit sampling and implementation safeguards and use the population quantities $p, y, z$ from Section 2, with $p \in (0, 1)$. Write $V(y) = y \odot (y - s_2\mathbf{1})$. With correct-mode bonuses $B_j^+ = g(1/(Ky_j) - 1)$ and zero bonus on incorrect modes,

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = \left(1 - \frac{g\varsigma(p)}{p}\right) V(y). \tag{A.1}$$

Thus $g$ is the advantage-bonus coefficient, while the effective gain in the collision field also depends on reward normalization. The base schedule $g = \lambda p/\varsigma(p)$ makes this effective gain equal to the configured parameter $\lambda$.

An optional common adjustment on incorrect modes preserves the instantaneous incorrect-mass drift at the same state. Its population form and the resulting correct-block field are

$$B^- = \frac{g(1/K - s_2)}{p(s_2 + t_2)}, \qquad \frac{\mathrm{d}y}{\mathrm{d}\tau} = \left[1 - \frac{g\varsigma(p)}{p}\frac{t_2 + 1/K}{s_2 + t_2}\right] V(y). \tag{A.2}$$

With this adjustment, the fully adaptive schedule

$$g = \lambda \frac{p}{\varsigma(p)} \frac{s_2 + t_2}{t_2 + 1/K} \tag{A.3}$$

again gives $\mathrm{d}y/\,\mathrm{d}\tau = (1 - \lambda)V(y)$: $\lambda = 1$ cancels within-correct competition and $\lambda > 1$ reverses it. These identities apply to the stated population update. Preserving the instantaneous incorrect-mass drift does not make the resulting $p(t)$ trajectories identical, since the conditional distributions change.

The practical method estimates the scheduling quantities within each rollout group and allows frequency floors, empirical centering, and bonus clipping. Configurations differ in their correction and scheduling settings; the 14B run at gain three also records an additional controller. The figure labels report the configured gain, not a measured coefficient of the neural training dynamics. The core construction and implementation safeguards are described by Rad et al. (2026a); no common field-reversal threshold is inferred for the plotted trajectories.

### A.2 Comparisons across regularization settings

The panel contains seven diversity-bonus runs, six covariance-based variants, two GRPO baselines with KL penalties, one entropy-bonus run, and one strong-KL run. Seven runs have nonzero diversity weights. Group size is 16 in thirteen runs and 8 in four; clustering settings also vary. These groups therefore do not form a controlled ablation of a single parameter. Figure 7 summarizes the within-run association between token entropy and collision. The Spearman coefficient is negative in 15 of 17 runs; the two positive values are $0.051$ and $0.008$. These associations complement the cutoff comparisons in the main text but do not establish that regularization preserves the unregularized flow or its limiting distribution.

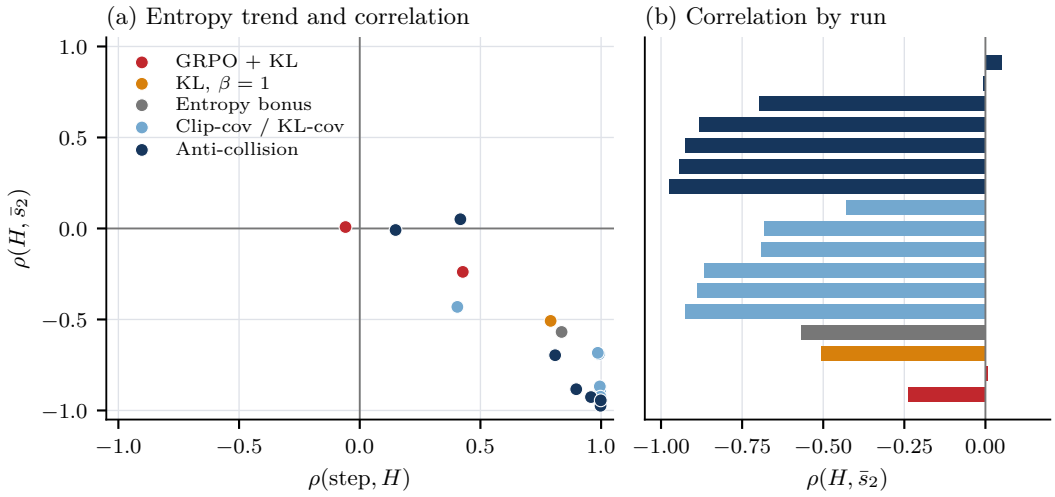


Figure 7: Descriptive entropy–collision associations across 17 DeepSeek-7B runs. Each point or bar summarizes one run under its logged training settings. Regime labels identify the main intervention; several anti-collision runs also use KL regularization. A negative association means higher logged token entropy co-occurs with lower correct-family collision. This association is not a prediction of the mode-level theorem for token entropy.

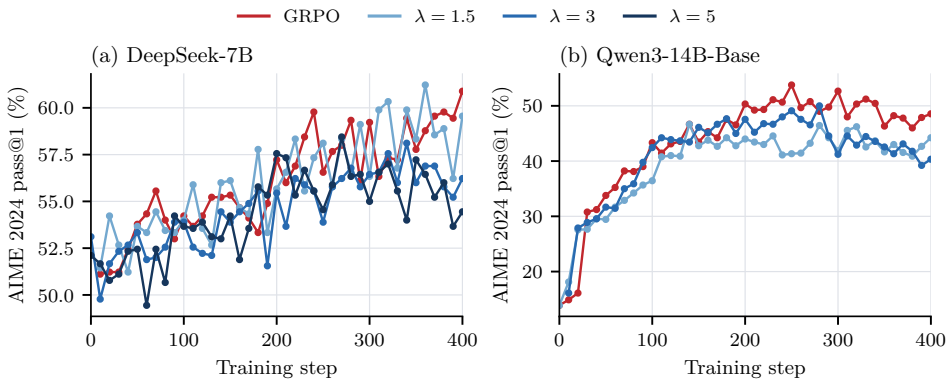


Figure 8: AIME 2024 validation accuracy under the tested anti-collision gains. Points are the logged mean@30 evaluations. Increasing family diversity does not consistently increase validation accuracy, particularly at the largest gain on Qwen3-14B-Base.

### A.3 Validation accuracy

Figure 8 reports the validation measurements available for the compared trajectories. The diversity changes in Figure 4 are not accompanied by a uniform accuracy ordering across gains. We therefore use the intervention to examine correct-family concentration and do not claim an accuracy improvement on both backbones.

## B Additional experimental details and supplementary plots

This appendix expands the experimental evidence in Section 7.3. The appendix panels use the same shifted-exponential and sigmoid forms as the main text, with each curve fit independently. RMSE values are reported in validation-accuracy percentage points. The source experiments are those of Cui et al. (2025); the math-task runs use DAPO-MATH (Yu et al., 2025), the "Zero" RL setting of Guo et al. (2025), the veRL/HybridFlow framework (Sheng et al., 2025), and GRPO-style updates related to Shao et al. (2024). Figures 9, 10, 11, and 12 show the per-panel fits used to produce Table 1.

Table 1: Finite-window surrogate comparison on 41 public curves from Cui et al. (2025). The row groups correspond to the following figures in Cui et al. (2025): base models from Figs. 3–4, early-prediction curves from Fig. 5, RL algorithms from Fig. 6, training-data curves from Fig. 13, and instruct-model curves from Fig. 14. RMSE is measured in validation-accuracy percentage points. AICc is used as a descriptive penalized-SSE summary; the goal is to compare which transformed bad-mass coordinate is closer to linear on each finite window, not to claim a universal replacement curve.

| Setting | Curves | Exp. RMSE | Sig. RMSE | $\Delta$ RMSE | Sig. AICc wins |
|---|---|---|---|---|---|
| All | 41 | 0.730 | 0.581 | 0.150 | 33/41 |
| Base models | 22 | 0.730 | 0.570 | 0.160 | 19/22 |
| Early prediction | 8 | 0.759 | 0.715 | 0.045 | 6/8 |
| RL algorithms | 4 | 0.655 | 0.574 | 0.080 | 3/4 |
| Training data | 4 | 0.951 | 0.480 | 0.471 | 3/4 |
| Instruct models | 3 | 0.462 | 0.443 | 0.019 | 2/3 |

**Fitting reproducibility.** The curve comparison uses 41 published entropy–accuracy traces from Cui et al. (2025). We fit the shifted exponential

$$R_{\exp}(H) = b - a\exp(H)$$

and the saturating sigmoid

$$R_\sigma(H) = R_{\min} + \frac{R_{\max} - R_{\min}}{1 + \exp\{\beta(H - H_0)\}},$$

fit independently to each public curve by bounded nonlinear least squares. AICc is computed as

$$\mathrm{AICc} = n\log(\mathrm{SSE}/n) + 2d + \frac{2d(d+1)}{n-d-1},$$

with $d = 2$ for the shifted exponential, $d = 3$ for the free-slope exponential, and $d = 4$ for the sigmoid; the same convention is used for BIC, $n\log(\mathrm{SSE}/n) + d\log n$. Here $d$ counts curve-shape parameters rather than a separate common residual-variance parameter, and $n$ is the number of retained points after sorting and duplicate removal, not necessarily the number of statistically independent RLVR checkpoints. Consequently, AICc and BIC should be interpreted as descriptive penalized residual scores rather than formal likelihood evidence: a dense, smooth published curve can make the effective sample size appear larger than the true number of independent checkpoints. The RMSE summaries, curve-level bootstrap in Appendix B.1, and held-out-by-checkpoint-order check are therefore the primary empirical summaries.

**Compute resources.** Curve fitting, analytic trajectory evaluation, and figure generation use CPU computation. The rollout comparisons use the logged statistics of Rad et al. (2026a); reproducing these summaries requires no additional model training. Recomputing family assignments requires the corresponding rollout texts and embedding model.

For derivative-based hazard diagnostics on public curves, we first sort each curve by entropy, average duplicate entropy locations, restrict to the stated accuracy window, and clip $\widehat{p}$ to $[10^{-6}, 1-10^{-6}]$ for numerical stability. We then fit shape-preserving cubic interpolants to the transformed coordinates $\mathrm{logit}\,\widehat{p}$ and $\log\widehat{p}$ as functions of entropy and evaluate their derivatives on a uniform grid inside the observed entropy interval. The reported coefficient of variation is $\mathrm{sd}(h)/|\mathrm{mean}(h)|$ over this grid. Exact synthetic trajectories use the analytic hazards from Appendix G. Because numerical differentiation of plots is sensitive to smoothing choices, these empirical hazard CVs are used as local descriptive diagnostics, not as standalone statistical tests.

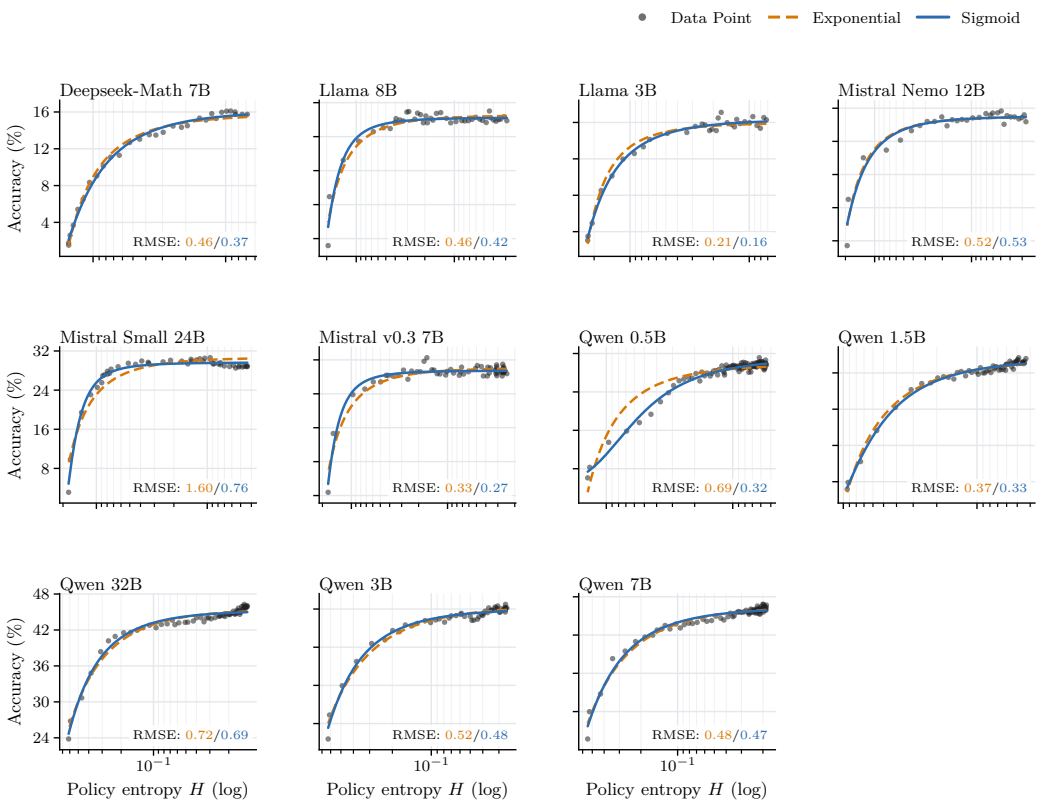


Figure 9: Math-task entropy–accuracy fits across base models. The sigmoid/logit surrogate is often closer on saturation windows than the shifted-exponential surrogate, especially in the mid-to-high accuracy regime where the curve saturates.

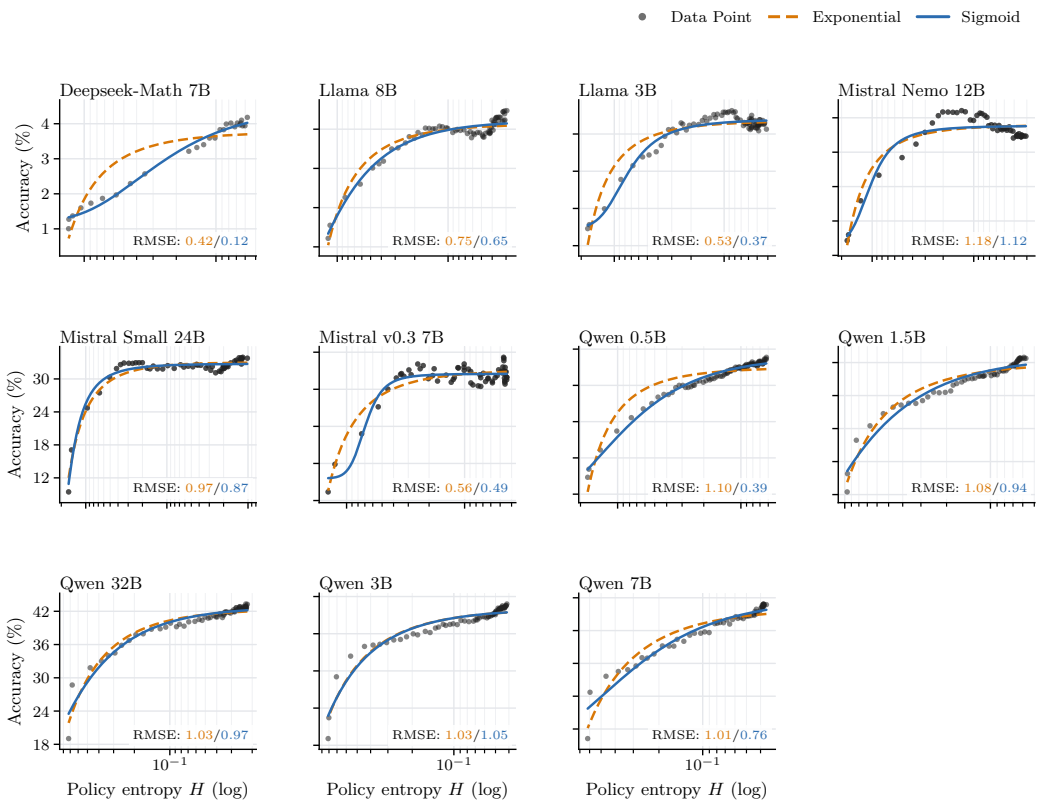


Figure 10: Code-task entropy–accuracy fits across base models. As in the math setting, the sigmoid/logit surrogate captures saturation effects that the shifted-exponential surrogate can miss.

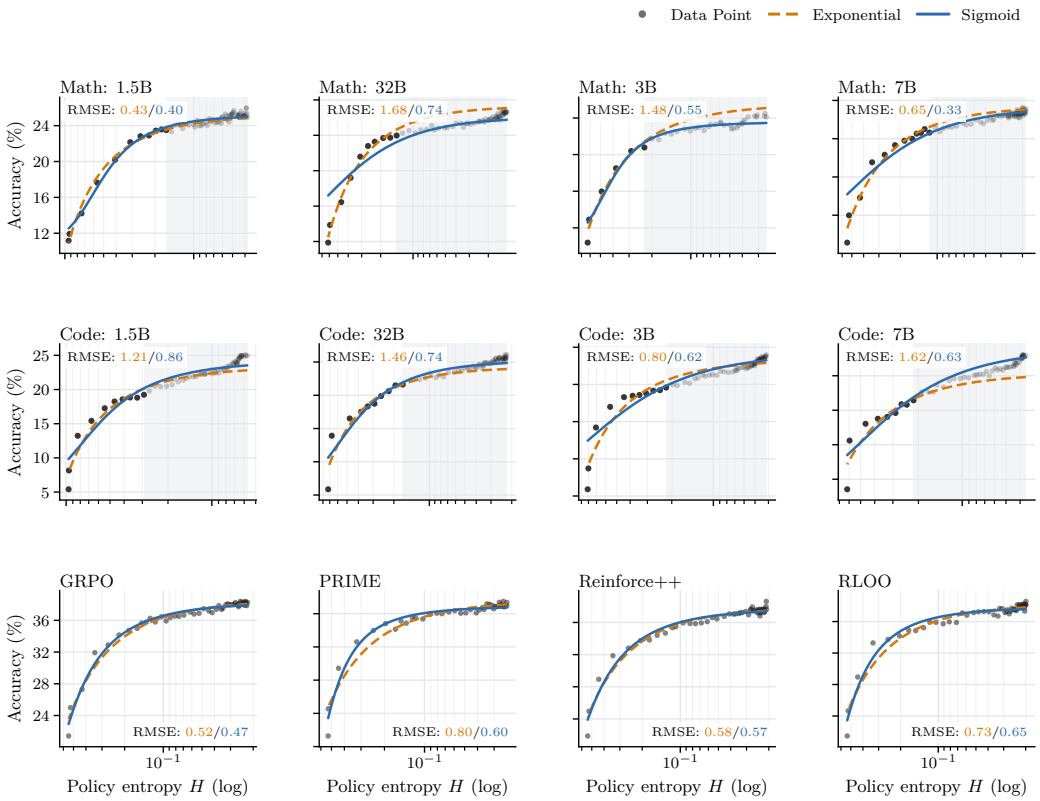


Figure 11: Early-prediction and algorithm ablations. The panels compare shifted-exponential and sigmoid fits for early-prediction curves and for Qwen2.5-7B trained with different RL algorithms.

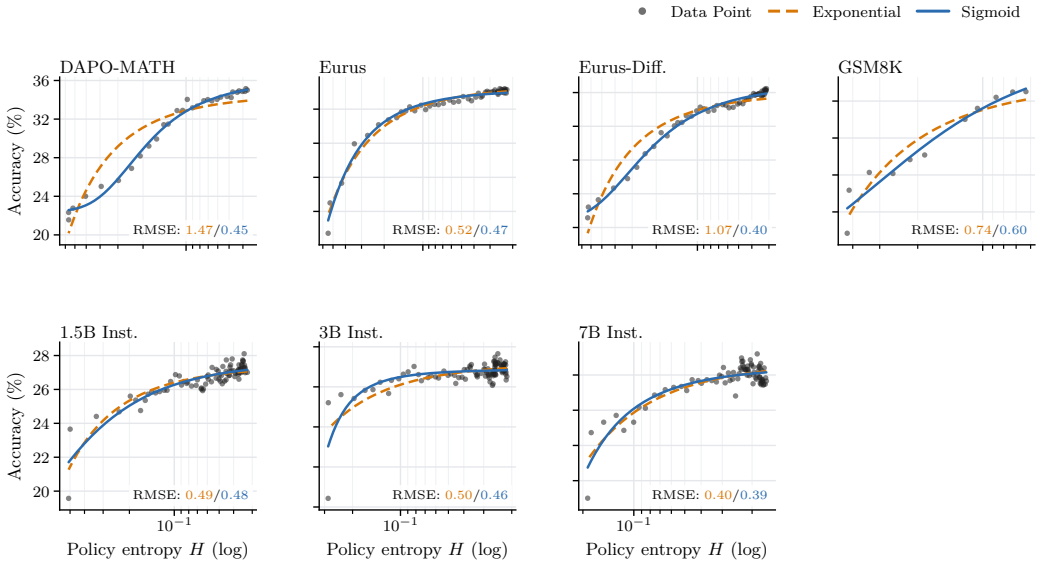


Figure 12: Training-data and instruct-model variants. The panels compare fits for Qwen2.5-7B trained with different data mixtures and for Qwen2.5 instruct models on the math task.

Table 2: Comparison of three finite-window curve families. RMSE is measured in validation-accuracy percentage points. The free-slope exponential $b - ae^{cH}$ has the lowest descriptive AICc score on most curves, while the sigmoid has the lowest mean RMSE.

| Model | Curves | Params | Mean RMSE | Pooled RMSE | AICc wins | BIC wins |
|---|---|---|---|---|---|---|
| Published exp. $b - ae^{H}$ | 41 | 2 | 0.730 | 0.796 | 4/41 | 6/41 |
| Free-slope exp. $b - ae^{cH}$ | 41 | 3 | 0.622 | 0.684 | 27/41 | 25/41 |
| Saturating sigmoid | 41 | 4 | 0.581 | 0.652 | 10/41 | 10/41 |

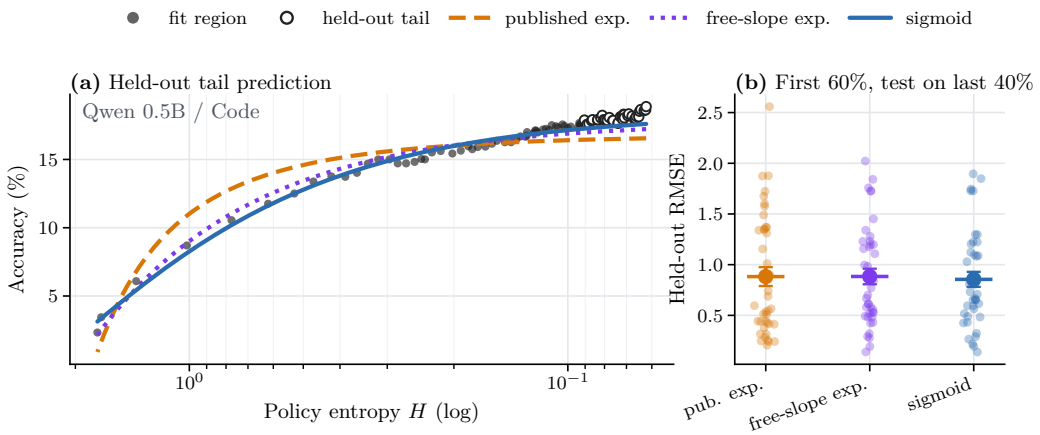


Figure 13: Held-out tail prediction check. Each curve is ordered by checkpoint index; models are fit on the first 60% of points and evaluated on the remaining 40%. The sigmoid has the lowest mean and pooled held-out RMSE, but the per-curve winners are mixed, consistent with the finite-window interpretation.

### B.1 Robustness to a free-slope exponential and held-out windows

The three-parameter free-slope exponential $b - a\exp(cH)$ provides a stronger baseline than the published unit-slope form. Table 2 shows that it improves substantially over the published form and is preferred more often by the descriptive AICc score. The sigmoid has the lowest mean and pooled RMSE. These results distinguish residual fit quality from parameter-penalized preference; they do not identify the latent prompt-level dynamics.

We summarize variation across curves by resampling whole trajectories rather than individual checkpoints. Positive RMSE differences favor the sigmoid. Resampling the 41 curves with replacement gives

$$\Delta_{\text{paper}-\text{exp}} = 0.150 \quad [0.084, 0.229], \qquad \Delta_{\text{free}-\text{exp}} = 0.041 \quad [0.007, 0.088],$$

where brackets denote 95% curve-resampling intervals. Curves from related source experiments need not be independent, so these intervals describe this collection rather than uncertainty across new training seeds. As a simple out-of-window check, fitting the first 60% of each trajectory and evaluating on the remaining 40% gives mean held-out RMSEs 0.882, 0.876, and 0.855 for the published exponential, free-slope exponential, and sigmoid, respectively. This check evaluates extrapolation along the reported trajectory order, with mixed per-curve winners.

### B.2 Hazard-flatness diagnostics

For the synthetic flatness atlas in Figure 14, we evaluate exact trajectories generated from five good-block initialization families: Dirichlet draws, one-head plus uniform tail, two-head plus uniform tail, one-head plus geometric tail, and one-head plus power-law tail.

$$\text{Dirichlet:} \qquad q \sim \mathrm{Dir}_K(\alpha \mathbf{1}), \qquad K \in \{4, 8, 16, 64\},\ \alpha \in \{0.1, 0.5, 1, 2\}; \tag{B.1}$$

$$\text{one-head + uniform tail:} \qquad q = \Big(w, \underbrace{\tfrac{1-w}{K-1}, \ldots, \tfrac{1-w}{K-1}}_{K-1 \text{ times}}\Big), \quad K \in \{4, 16, 64, 256\}; \tag{B.2}$$

$$\text{two-head + uniform tail:} \qquad q = \Big(w_1, w_2, \underbrace{\tfrac{1-w_1-w_2}{K-2}, \ldots, \tfrac{1-w_1-w_2}{K-2}}_{K-2 \text{ times}}\Big), \quad K \in \{8, 32, 128\}; \tag{B.3}$$

$$\text{one-head + geometric tail:} \qquad q_1 = w, \quad q_j = (1-w)\frac{r^{j-2}}{\sum_{m=0}^{K-2} r^m}, \quad j \geq 2, \quad K \in \{16, 64, 256\}; \tag{B.4}$$

$$\text{one-head + power-law tail:} \qquad q_1 = w, \quad q_j = (1-w)\frac{(j-1)^{-\rho}}{\sum_{m=1}^{K-1} m^{-\rho}}, \quad j \geq 2, \quad K \in \{16, 64, 256\}. \tag{B.5}$$

The exact trajectories use a single active bad arm for the summary in Table 3. The 396 synthetic trajectories are generated as follows: the Dirichlet family uses 10 independent draws for each $(K, \alpha)$ pair with random seed 0; the one-head uniform family uses 20 evenly spaced values of $w \in [0.05, 0.95]$ for each listed $K$; the two-head family uses the grid $w_1 \in \{0.25, 0.40, 0.55, 0.70\}$, $w_2 \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ for each listed $K$; the geometric family uses $w \in \{0.2, 0.4, 0.6, 0.8\}$ and $r \in \{0.3, 0.5, 0.7, 0.9\}$; and the power-law family uses $w \in \{0.2, 0.4, 0.6, 0.8\}$ and tail exponent $\rho \in \{0.5, 1.0, 1.5, 2.0\}$. For each trajectory or curve, we compute the coefficient of variation of the logit-side hazard $\Lambda$ and the residual-exponential hazard $\mathcal{B}$, along with the transformed-coordinate linearity scores $R^2_{\mathrm{logit}}$ and $R^2_{\mathrm{exp}}$. Figure 14 provides the corresponding detailed hazard plots.

Table 3: Exact hazard-flatness diagnostics in the concentrated-bad-block simulation library. A flat $\Lambda(H) = \mathrm{d}\,\mathrm{logit}(p)/\,\mathrm{d}H$ implies a logistic bridge, while a flat $\mathcal{B}(H) = \mathrm{d}\log p/\,\mathrm{d}H$ implies the shifted-exponential residual bridge.

| Accuracy window | Curves | Mean CV($\Lambda$) | Mean CV($\mathcal{B}$) | Mean $R^2_{\mathrm{logit}}$ | Mean $R^2_{\mathrm{exp}}$ |
|---|---|---|---|---|---|
| 5–45% | 396 | 0.135 | 0.208 | 0.995 | 0.988 |
| 45–75% | 396 | 0.071 | 0.220 | 0.999 | 0.989 |
| 75–95% | 396 | 0.111 | 0.404 | 0.997 | 0.963 |
| 95–99.5% | 396 | 0.171 | 0.520 | 0.993 | 0.939 |

### B.3 Direct fits of the terminal entropy asymptotic

To make the terminal asymptotic more concrete, Figure 15 fits controlled exact trajectories that satisfy the unique-winner assumption used in Appendix D. For each uniform bad block size $M$, the structural exponent is

$$\alpha = \frac{M}{M+1} \in [1/2, 1).$$

On the terminal tail we regress

$$\log\big(H(p)/|\log p|\big) = \log C + \widehat{\alpha} \log p$$

on the range $10^{-12} \leq p \leq 10^{-7}$. This is a check of the asymptotic prediction $H(p) \sim Cp^{\alpha}|\log p|$, not a free finite-window exponent. Broad-window fits on trajectories with tied or nearly tied good maxima can produce effective slopes below $1/2$; those slopes are preasymptotic diagnostics and should not be interpreted as the structural $\alpha$.

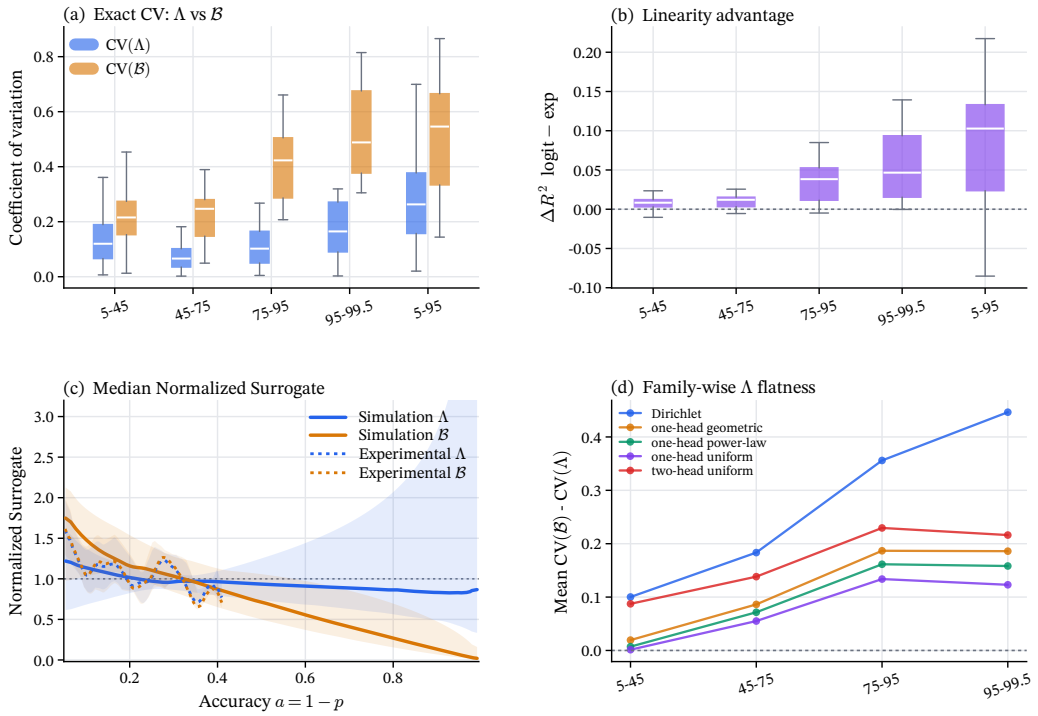


Figure 14: Additional exact and empirical hazard diagnostics. The exact trajectories illustrate intervals with relatively stable logit-side slopes. Derivative estimates from digitized aggregate curves are descriptive and sensitive to interpolation. The separate comparison in Section 7.3 evaluates the resulting curve families.

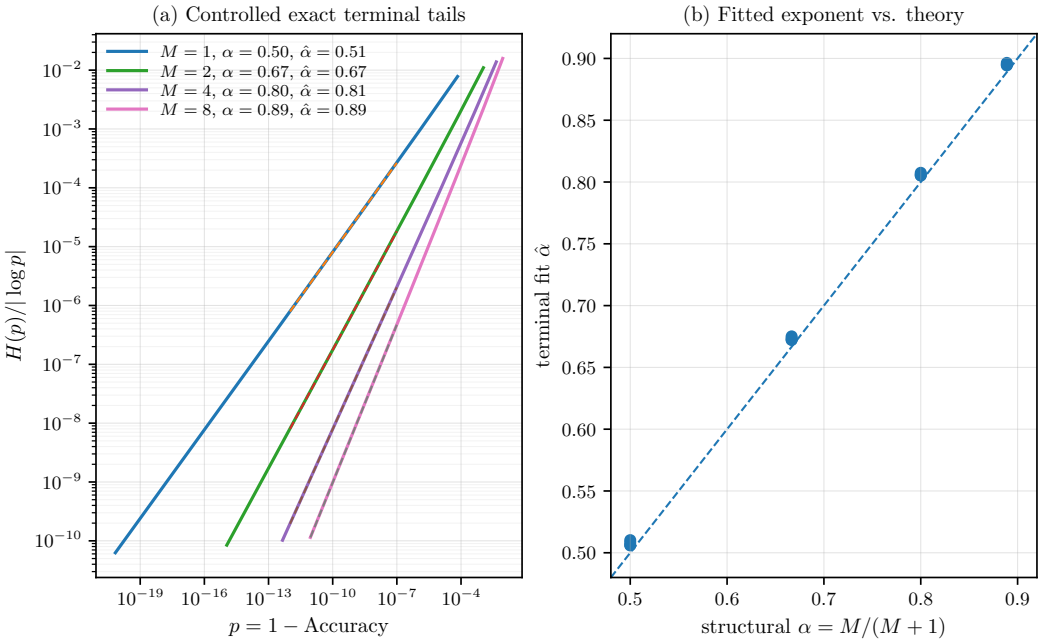


Figure 15: Controlled direct check of the terminal $p^{\alpha}|\log p|$ asymptotic. All plotted trajectories have a unique, separated good-block maximizer and a uniform bad block of size $M$. Left: representative exact tails plotted as $H(p)/|\log p|$ against $p$ on log–log axes, with dashed fitted lines over $10^{-12} \leq p \leq 10^{-7}$. Right: fitted terminal exponents versus the structural prediction $\alpha = M/(M+1)$ across four unique-winner good-block initializations and $M \in \{1, 2, 4, 8\}$; the dashed line is equality. For $M = 1$, the true terminal exponent is exactly $1/2$. Values below $1/2$ arise only as finite-window or assumption-violating effective slopes, not as structural terminal exponents.

# C SCALAR REDUCTION AND EXACT ENTROPY IDENTITIES

The good–bad collision flow admits an exact scalar parameterization. We use the reciprocal transformation of Rad et al. (2026b) to solve each conditional block, match their shared internal time, and derive the entropy and mass identities below. Throughout these appendices, $H$ denotes the good-block Shannon entropy $H_y$; it is distinct from the collision entropy $-\log s_2$ and from a logged token-entropy statistic.

## C.1 SETUP

Let

$$y(\tau) \in \Delta^{K-1}, \qquad z(\tau) \in \Delta^{M-1}, \qquad p(\tau) \in (0,1),$$

denote the normalized good-block composition, the normalized bad-block composition, and the total bad mass. Unless a statement explicitly includes zero coordinates, $K \geq 1$ and $M \geq 1$ count the active initial supports, so the initial block distributions are strictly positive and $p_0 \in (0,1)$. Families are fixed throughout the trajectory. In internal time $\tau$, the dynamics are

$$\frac{dy_j}{d\tau} = y_j\big(y_j - s_2\big), \qquad s_2 := \sum_{i=1}^{K} y_i^2, \tag{C.1}$$

$$\frac{dz_m}{d\tau} = z_m\big(t_2 - z_m\big), \qquad t_2 := \sum_{\ell=1}^{M} z_\ell^2, \tag{C.2}$$

$$\frac{dp}{d\tau} = -p(1-p)\big(s_2 + t_2\big). \tag{C.3}$$

We write

$$q := y(0) \in \Delta^{K-1}, \qquad r := z(0) \in \Delta^{M-1}, \qquad q_{\max} := \max_j q_j.$$

**Lemma C.1** (Simplex invariance)**.** *If $y(0) \in \Delta^{K-1}$ and $z(0) \in \Delta^{M-1}$, then $y(\tau) \in \Delta^{K-1}$ and $z(\tau) \in \Delta^{M-1}$ for all $\tau \geq 0$. Moreover, zero coordinates remain zero.*

*Proof.* Let $Y(\tau) = \sum_j y_j(\tau)$ and $Z(\tau) = \sum_m z_m(\tau)$. The coordinate equations give

$$Y' = s_2(1 - Y), \qquad Z' = t_2(Z - 1).$$

With $Y(0) = Z(0) = 1$, uniqueness for these scalar linear equations gives $Y = Z = 1$. Each coordinate equation has the form $x' = xf(\tau)$, so a positive coordinate remains positive at every finite time and a zero coordinate remains zero. The trajectory therefore stays in the compact product of simplices, which also ensures existence for all $\tau \geq 0$. □

## C.2 GOOD-BLOCK SCALAR REDUCTION

**Lemma C.2** (Reciprocal linearization for the good block)**.** *For every index $j$ with $q_j > 0$, define*

$$u_j(\tau) := \frac{1}{y_j(\tau)}.$$

*Then $u_j$ solves the linear ODE*

$$u_j' - s_2 u_j = -1. \tag{C.4}$$

*Proof.* Differentiating $u_j = 1/y_j$ and using (C.1),

$$u_j' = -\frac{y_j'}{y_j^2} = -\frac{y_j(y_j - s_2)}{y_j^2} = -1 + s_2 u_j.$$

This is exactly (C.4). □

**Proposition C.3** (Good-block scalar reduction)**.** *Define*

$$S(\tau) := \int_0^\tau s_2(u)\,du, \qquad I(\tau) := \int_0^\tau e^{-S(u)}\,du,$$

*and*

$$A(I) := \sum_{j=1}^K \frac{q_j}{1 - Iq_j}, \qquad B(I) := \sum_{j=1}^K \frac{q_j^2}{(1 - Iq_j)^2}.$$

*Then*

$$y_j(\tau) = y_j\big(I(\tau)\big) = \frac{\dfrac{q_j}{1 - I(\tau)q_j}}{A\big(I(\tau)\big)}, \qquad I \in [0, 1/q_{\max}), \tag{C.5}$$

*and*

$$\frac{dI}{d\tau} = e^{-S(\tau)} = \frac{1}{A\big(I(\tau)\big)} > 0, \qquad \frac{d\tau}{dI} = A(I). \tag{C.6}$$

*Consequently,*

$$\tau(I) = \int_0^I A(u)\,du = -\sum_{j=1}^K \log(1 - Iq_j), \tag{C.7}$$

*and*

$$e^{-\tau(I)} = \prod_{j=1}^K (1 - Iq_j). \tag{C.8}$$

*If $q$ has a unique maximizer $q_{j_*} = q_{\max}$, then*

$$y_{j_*}(\tau) \to 1, \qquad y_j(\tau) \to 0 \ \ (j \neq j_*), \qquad I(\tau) \uparrow \frac{1}{q_{\max}} \quad (\tau \to \infty). \tag{C.9}$$

*Proof.* Fix $j$ with $q_j > 0$. By Lemma C.2, $u_j$ solves

$$u_j' - s_2 u_j = -1.$$

Multiplying by the integrating factor $e^{-S(\tau)}$ gives

$$\frac{d}{d\tau}\big(e^{-S(\tau)} u_j(\tau)\big) = -e^{-S(\tau)}.$$

Integrating from $0$ to $\tau$,

$$e^{-S(\tau)} u_j(\tau) - u_j(0) = -\int_0^\tau e^{-S(s)}\,ds = -I(\tau).$$

Since $u_j(0) = 1/q_j$,

$$u_j(\tau) = e^{S(\tau)}\Big(\frac{1}{q_j} - I(\tau)\Big).$$

Hence

$$y_j(\tau) = \frac{1}{u_j(\tau)} = e^{-S(\tau)} \frac{q_j}{1 - I(\tau)q_j}.$$

Summing over $j$ and using $\sum_j y_j(\tau) = 1$,

$$1 = e^{-S(\tau)} \sum_{j=1}^K \frac{q_j}{1 - I(\tau)q_j} = e^{-S(\tau)} A\big(I(\tau)\big),$$

so

$$e^{-S(\tau)} = \frac{1}{A\big(I(\tau)\big)}.$$

Substituting this into the previous display yields (C.5). Differentiating the definition of $I$ gives (C.6), and integrating $d\tau/dI = A(I)$ gives

$$\tau(I) = \sum_{j=1}^{K} \int_0^I \frac{q_j}{1 - uq_j}\, du = -\sum_{j=1}^{K} \log(1 - Iq_j),$$

which proves (C.7); exponentiating gives (C.8).

The admissible interval is $I < 1/q_{\max}$, since every denominator $1 - Iq_j$ with $q_j > 0$ must remain positive. If $q$ has a unique maximizer $q_{j_*} = q_{\max}$, then as $I \uparrow 1/q_{\max}$, the $j_*$-term in (C.5) diverges while every other term stays finite. Therefore the normalized vector converges to $e_{j_*}$. Since (C.7) contains the divergent term $-\log(1 - Iq_{\max})$, one also has $\tau(I) \to \infty$ as $I \uparrow 1/q_{\max}$. □

*Remark* C.4 (Ties at the maximum). If the maximizer is not unique, let

$$\mathcal{M} := \{j : q_j = q_{\max}\}, \qquad |\mathcal{M}| = m > 1.$$

Then

$$y_j(\tau) \to \begin{cases} 1/m, & j \in \mathcal{M}, \\ 0, & j \notin \mathcal{M}, \end{cases} \qquad \tau \to \infty.$$

Indeed, all divergent terms in (C.5) have the same leading order $q_{\max}/(1 - Iq_{\max})$, so the mass is shared uniformly across the tie set.

### C.3 Bad-block scalar reduction

**Lemma C.5** (Reciprocal linearization for the bad block)**.** *For every index $m$ with $r_m > 0$, define*

$$v_m(\tau) := \frac{1}{z_m(\tau)}.$$

*Then $v_m$ solves the linear ODE*

$$v_m' + t_2 v_m = 1. \tag{C.10}$$

*Proof.* Differentiating $v_m = 1/z_m$ and using (C.2),

$$v_m' = -\frac{z_m'}{z_m^2} = -\frac{z_m(t_2 - z_m)}{z_m^2} = 1 - t_2 v_m.$$

This is (C.10). □

**Proposition C.6** (Bad-block scalar reduction)**.** *Define*

$$T(\tau) := \int_0^\tau t_2(u)\, du, \qquad U(\tau) := \int_0^\tau e^{T(u)}\, du,$$

*and*

$$A_-(U) := \sum_{m=1}^{M} \frac{r_m}{1 + Ur_m}, \qquad B_-(U) := \sum_{m=1}^{M} \frac{r_m^2}{(1 + Ur_m)^2}.$$

*Then*

$$z_m(\tau) = z_m\big(U(\tau)\big) = \frac{r_m}{1 + U(\tau) r_m} \Big/ A_-\big(U(\tau)\big), \qquad U \in [0, \infty), \tag{C.11}$$

*and*

$$\frac{dU}{d\tau} = e^{T(\tau)} = \frac{1}{A_-\big(U(\tau)\big)} > 0, \qquad \frac{d\tau}{dU} = A_-(U). \tag{C.12}$$

*Consequently,*

$$\tau(U) = \int_0^U A_-(u)\, du = \sum_{m=1}^{M} \log(1 + Ur_m), \tag{C.13}$$

*and*

$$e^{\tau(U)} = \prod_{m=1}^{M} (1 + Ur_m), \qquad e^{-T(U)} = A_-(U). \tag{C.14}$$

*Finally,*

$$U(\tau) \uparrow \infty, \qquad z(\tau) \to \frac{1}{M}\mathbf{1} \qquad (\tau \to \infty). \tag{C.15}$$

*Proof.* Fix $m$ with $r_m > 0$. By Lemma C.5,

$$v_m' + t_2 v_m = 1.$$

Multiplying by the integrating factor $e^{T(\tau)}$,

$$\frac{d}{d\tau}\big(e^{T(\tau)} v_m(\tau)\big) = e^{T(\tau)}.$$

Integrating from 0 to $\tau$,

$$e^{T(\tau)} v_m(\tau) - v_m(0) = \int_0^\tau e^{T(s)}\, ds = U(\tau).$$

Since $v_m(0) = 1/r_m$,

$$v_m(\tau) = e^{-T(\tau)}\Big(\frac{1}{r_m} + U(\tau)\Big).$$

Therefore

$$z_m(\tau) = \frac{1}{v_m(\tau)} = e^{T(\tau)} \frac{r_m}{1 + U(\tau) r_m}.$$

Summing over $m$ and using $\sum_m z_m(\tau) = 1$,

$$1 = e^{T(\tau)} \sum_{m=1}^{M} \frac{r_m}{1 + U(\tau) r_m} = e^{T(\tau)} A_-\big(U(\tau)\big),$$

so

$$e^{T(\tau)} = \frac{1}{A_-\big(U(\tau)\big)}.$$

Substituting back yields (C.11). The relations (C.12) and (C.13) follow exactly as in the good-block case:

$$\tau(U) = \sum_{m=1}^{M} \int_0^U \frac{r_m}{1 + u r_m}\, du = \sum_{m=1}^{M} \log(1 + U r_m).$$

Exponentiating proves (C.14).

As $U \to \infty$,

$$\frac{r_m}{1 + U r_m} \sim \frac{1}{U} \qquad \text{for all } m,$$

hence

$$A_-(U) \sim \frac{M}{U}, \qquad z_m(U) \sim \frac{1/U}{M/U} = \frac{1}{M}.$$

This proves (C.15). □

**Lemma C.7** (Coupled one-parameter representation)**.** *For every admissible $I \in [0, 1/q_{\max})$, there exists a unique $U(I) \in [0, \infty)$ such that*

$$-\sum_{j=1}^{K} \log(1 - I q_j) = \sum_{m=1}^{M} \log(1 + U(I) r_m). \tag{C.16}$$

*Equivalently,*

$$\tau(I) = \tau\big(U(I)\big).$$

*Moreover,*

$$\frac{dU}{dI} = \frac{A(I)}{A_-\big(U(I)\big)}. \tag{C.17}$$

*Proof.* By Propositions C.3 and C.6, both $\tau(I)$ and $\tau(U)$ are continuous and strictly increasing, with

$$\tau(I) \uparrow \infty \quad \text{as } I \uparrow 1/q_{\max}, \qquad \tau(U) \uparrow \infty \quad \text{as } U \uparrow \infty.$$

Therefore, for every admissible $I$, there exists a unique $U(I)$ such that $\tau(I) = \tau(U(I))$. Differentiating this identity yields

$$A(I) = A_-\big(U(I)\big)\frac{dU}{dI},$$

which is (C.17). □

### C.4 Entropy-collapse theorem for the collision flow

The scalar parameterization yields the concentration and entropy limits stated in the main text. The following theorem concerns the finite family distribution governed by the collision field of Rad et al. (2026b).

For this subsection, restrict to the active supports of the initial condition; equivalently, after discarding zero-mass modes, assume that $q_j > 0$ for all $j \in [K]$ and $r_m > 0$ for all $m \in [M]$. Define

$$H_y(\tau) := -\sum_{j=1}^{K} y_j(\tau)\log y_j(\tau), \qquad H_z(\tau) := -\sum_{m=1}^{M} z_m(\tau)\log z_m(\tau),$$

and recall $s_2 = \sum_j y_j^2$, $t_2 = \sum_m z_m^2$. The Shannon entropy of the full family distribution is

$$H_{\mathrm{mode}}(\tau) = -\big(1-p(\tau)\big)\log\big(1-p(\tau)\big) - p(\tau)\log p(\tau) + \big(1-p(\tau)\big)H_y(\tau) + p(\tau)H_z(\tau). \tag{C.18}$$

*Theorem* C.1 (Mode-level entropy collapse under the GRPO collision field). Consider the internal-time dynamics (C.1)–(C.3) with active-support sizes $K$ and $M$. Then the following hold.

(i) **Good modes concentrate; bad modes spread.**

$$\frac{ds_2}{d\tau} = 2\sum_{j=1}^{K} y_j\big(y_j - s_2\big)^2 \geq 0, \qquad \frac{dt_2}{d\tau} = -2\sum_{m=1}^{M} z_m\big(z_m - t_2\big)^2 \leq 0. \tag{C.19}$$

Equivalently, the good-block Rényi-2 entropy $-\log s_2$ decreases, while the bad-block Rényi-2 entropy $-\log t_2$ increases.

(ii) **Shannon entropy obeys the same signed law.**

$$\frac{dH_y}{d\tau} = -\operatorname{Cov}_{j\sim y}\big(y_j, \log y_j\big) \leq 0, \qquad \frac{dH_z}{d\tau} = \operatorname{Cov}_{m\sim z}\big(z_m, \log z_m\big) \geq 0. \tag{C.20}$$

Equality holds in either display iff the corresponding block is uniform on its active support.

(iii) **Rank order is preserved and the initial leader selects the terminal face.** For positive coordinates,

$$\frac{d}{d\tau}\log\frac{y_i}{y_j} = y_i - y_j, \qquad \frac{d}{d\tau}\log\frac{z_m}{z_\ell} = -(z_m - z_\ell). \tag{C.21}$$

Thus equality hyperplanes are invariant. If $\mathcal{M} := \{j : q_j = q_{\max}\}$ has size $m_\star$, then

$$y_j(\tau) \longrightarrow \begin{cases} 1/m_\star, & j \in \mathcal{M}, \\ 0, & j \notin \mathcal{M}, \end{cases} \qquad H_y(\tau) \longrightarrow \log m_\star. \tag{C.22}$$

In particular, a unique initial maximizer gives $y(\tau) \to e_{j_\star}$ and $H_y(\tau) \to 0$.

(iv) **Bad mass vanishes and the family distribution has a finite-support limit.** The bad-mass logit satisfies

$$\frac{d}{d\tau}\operatorname{logit} p(\tau) = -(s_2(\tau) + t_2(\tau)). \tag{C.23}$$

Consequently $p(\tau) \to 0$. Since $z(\tau) \to M^{-1}\mathbf{1}$, one has

$$H_{\mathrm{mode}}(\tau) \longrightarrow \log m_\star. \tag{C.24}$$

Under a unique initial good-family leader, the family entropy tends to zero. This limit concerns family selection; it does not require token strings within the selected family to become deterministic.

*Proof.* For $s_2$, differentiate and use (C.1):

$$\frac{ds_2}{d\tau} = 2\sum_j y_j \frac{dy_j}{d\tau} = 2\sum_j y_j^2(y_j - s_2) = 2\Big(\sum_j y_j^3 - s_2^2\Big) = 2\sum_j y_j(y_j - s_2)^2.$$

The last expression is nonnegative. The same computation with (C.2) gives

$$\frac{dt_2}{d\tau} = 2\sum_m z_m^2(t_2 - z_m) = -2\sum_m z_m(z_m - t_2)^2 \le 0.$$

This proves (C.19); the Rényi-2 statement follows by differentiating $-\log s_2$ and $-\log t_2$.

For Shannon entropy, $\sum_j y_j' = 0$, so

$$\frac{dH_y}{d\tau} = -\sum_j y_j' \log y_j = -\sum_j y_j(y_j - s_2)\log y_j = -\operatorname{Cov}_{j\sim y}(y_j, \log y_j).$$

The covariance is nonnegative because $a \mapsto a$ and $a \mapsto \log a$ are increasing functions on $(0, 1]$. More explicitly,

$$\operatorname{Cov}_{j\sim y}(y_j, \log y_j) = \frac{1}{2}\sum_{i,j} y_i y_j\,(y_i - y_j)\big(\log y_i - \log y_j\big) \ge 0.$$

Hence $dH_y/d\tau \le 0$, with equality iff all positive $y_j$'s are equal. Repeating the same calculation for $z_m' = z_m(t_2 - z_m)$ gives

$$\frac{dH_z}{d\tau} = -\sum_m z_m(t_2 - z_m)\log z_m = \operatorname{Cov}_{m\sim z}(z_m, \log z_m) \ge 0,$$

again with equality iff $z$ is uniform on its active support. This proves (C.20).

For rank preservation, divide the coordinate equations by the corresponding positive coordinate:

$$\frac{d}{d\tau}\log y_j = y_j - s_2, \qquad \frac{d}{d\tau}\log z_m = t_2 - z_m.$$

Subtracting two such identities gives (C.21). Therefore equality hyperplanes cannot be crossed. The terminal statement for $y$ follows from the scalar representation (C.5): as $I \uparrow 1/q_{\max}$, exactly the terms with $q_j = q_{\max}$ diverge, all with the same leading order, while all other terms remain finite. Normalizing gives (C.22). The bad-block scalar representation (C.11) similarly gives $z(\tau) \to M^{-1}\mathbf{1}$.

Finally, (C.3) implies (C.23). Since

$$\frac{1}{K} \le s_2(\tau) \le 1, \qquad \frac{1}{M} \le t_2(\tau) \le 1,$$

we have

$$\operatorname{logit} p(\tau) \le \operatorname{logit} p(0) - \Big(\frac{1}{K} + \frac{1}{M}\Big)\tau \longrightarrow -\infty,$$

and therefore $p(\tau) \to 0$. Taking the limit in (C.18), using the boundedness of $H_z$ and the binary entropy term, yields (C.24). If $m_\star = 1$, this limit is zero. □

*Remark* C.8 (Conditional and total entropy). Good-family Shannon and collision entropies are nonincreasing, while the corresponding bad-family entropies are nondecreasing. The full family entropy need not be monotone, even though its terminal value is fixed by the initial good-family maximizers. For example, when $K = M = 1$ and $p_0 > 1/2$, it equals $h_{\mathrm{bin}}(p)$ and initially increases as $p$ decreases toward $1/2$. Token-level conclusions in Appendix F additionally assume that each arm represents one complete token string.

### C.5 Exact entropy and bad-mass formulas

We define

$$A(I) := \sum_{j=1}^{K} \frac{q_j}{1 - Iq_j}, \qquad B(I) := \sum_{j=1}^{K} \frac{q_j^2}{(1 - Iq_j)^2},$$

$$C_q(I) := \sum_{j=1}^{K} \frac{q_j \log q_j}{1 - Iq_j}, \qquad C_1(I) := \sum_{j=1}^{K} \frac{q_j \log(1 - Iq_j)}{1 - Iq_j},$$

$$D_q(I) := \sum_{j=1}^{K} \frac{q_j^2 \log q_j}{(1 - Iq_j)^2}, \qquad D_1(I) := \sum_{j=1}^{K} \frac{q_j^2 \log(1 - Iq_j)}{(1 - Iq_j)^2},$$

and, for the bad block,

$$A_-(U) := \sum_{m=1}^{M} \frac{r_m}{1 + Ur_m}, \qquad B_-(U) := \sum_{m=1}^{M} \frac{r_m^2}{(1 + Ur_m)^2}.$$

**Proposition C.9** (Good-block entropy, collision, and bad-mass identities)**.** *Let $U(I)$ be given by Lemma C.7, and define*

$$\mathcal{S}(I) := \sum_{j=1}^{K} y_j(I)^2 = \frac{B(I)}{A(I)^2}, \qquad \mathcal{T}(I) := t_2(\tau(I)) = \frac{B_-\big(U(I)\big)}{A_-\big(U(I)\big)^2}, \tag{C.25}$$

*and*

$$\mathcal{C}(I) := \mathrm{Cov}_{j \sim y(I)}\big(y_j, \log y_j\big) = \frac{D_q(I) - D_1(I)}{A(I)^2} + \frac{B(I)}{A(I)^3}\big(C_1(I) - C_q(I)\big). \tag{C.26}$$

*Then*

$$H(I) := -\sum_{j=1}^{K} y_j(I) \log y_j(I) = \log A(I) + \frac{C_1(I) - C_q(I)}{A(I)}, \tag{C.27}$$

$$\frac{dH}{dI} = -A(I)\mathcal{C}(I), \tag{C.28}$$

*and*

$$\operatorname{logit} p(I) = \operatorname{logit} p_0 - \log A(I) + \log A_-\big(U(I)\big). \tag{C.29}$$

*Equivalently,*

$$\frac{p(I)}{1 - p(I)} = \frac{p_0}{1 - p_0} \, \frac{A_-\big(U(I)\big)}{A(I)}. \tag{C.30}$$

*In particular,*

$$\frac{d}{dI} \operatorname{logit} p(I) = -\big(\mathcal{S}(I) + \mathcal{T}(I)\big)A(I), \tag{C.31}$$

*so $I \mapsto p(I)$ is strictly decreasing, and*

$$\frac{dH}{dp} = \frac{\mathcal{C}\big(I(p)\big)}{p(1-p)\big(\mathcal{S}\big(I(p)\big) + \mathcal{T}\big(I(p)\big)\big)}. \tag{C.32}$$

*Proof.* The formula for $\mathcal{S}(I)$ follows directly from (C.5):

$$\sum_j y_j(I)^2 = \sum_j \frac{\dfrac{q_j^2}{(1 - Iq_j)^2}}{A(I)^2} = \frac{B(I)}{A(I)^2}.$$

Likewise,

$$\mathcal{T}(I) = \sum_m z_m \big(U(I)\big)^2 = \frac{B_-\big(U(I)\big)}{A_-\big(U(I)\big)^2}.$$

Next, from (C.5),

$$\log y_j(I) = \log q_j - \log(1 - Iq_j) - \log A(I),$$

hence

$$H(I) = -\sum_j y_j \log y_j = \log A(I) - \sum_j y_j \log q_j + \sum_j y_j \log(1 - Iq_j).$$

Using the explicit form of $y_j(I)$,

$$\sum_j y_j \log q_j = \frac{C_q(I)}{A(I)}, \qquad \sum_j y_j \log(1 - Iq_j) = \frac{C_1(I)}{A(I)},$$

which proves (C.27).

For the derivative, differentiate $H(I) = -\sum_j y_j \log y_j$:

$$\frac{dH}{dI} = -\sum_j y_j'(I)\big(1 + \log y_j(I)\big).$$

Since $\sum_j y_j'(I) = 0$, this simplifies to

$$\frac{dH}{dI} = -\sum_j y_j'(I) \log y_j(I).$$

Now

$$\frac{dy_j}{dI} = \frac{dy_j/d\tau}{dI/d\tau} = A(I)\, y_j(I)\big(y_j(I) - \mathcal{S}(I)\big),$$

because $dI/d\tau = 1/A(I)$. Therefore

$$\frac{dH}{dI} = -A(I) \sum_j y_j(I)\big(y_j(I) - \mathcal{S}(I)\big) \log y_j(I) = -A(I)\mathcal{C}(I),$$

which is (C.28). The explicit representation (C.26) is the corresponding covariance expanded in terms of the scalar sums.

For the bad mass, (C.3) gives

$$\frac{d}{d\tau} \operatorname{logit} p = -(s_2 + t_2).$$

Integrating from 0 to $\tau$,

$$\operatorname{logit} p(\tau) = \operatorname{logit} p_0 - \int_0^\tau s_2(u)\, du - \int_0^\tau t_2(u)\, du = \operatorname{logit} p_0 - S(\tau) - T(\tau).$$

By Proposition C.3,

$$S(\tau(I)) = \log A(I),$$

and by Proposition C.6,

$$T\big(\tau(U)\big) = -\log A_-(U).$$

Thus

$$\operatorname{logit} p(I) = \operatorname{logit} p_0 - \log A(I) + \log A_-\big(U(I)\big),$$

which proves (C.29); exponentiating gives (C.30).

Differentiating (C.29) and using $dU/dI = A(I)/A_-(U(I))$,

$$\frac{d}{dI} \operatorname{logit} p(I) = -\frac{B(I)}{A(I)} - \frac{B_-\big(U(I)\big)}{A_-\big(U(I)\big)} \frac{A(I)}{A_-\big(U(I)\big)} = -\big(\mathcal{S}(I) + \mathcal{T}(I)\big)A(I),$$

which is (C.31). Hence

$$\frac{dp}{dI} = p(1-p)\frac{d}{dI} \operatorname{logit} p(I) = -p(1-p)\big(\mathcal{S}(I) + \mathcal{T}(I)\big)A(I) < 0.$$

Dividing (C.28) by this identity yields (C.32). □

*Remark* C.10 (Single active bad arm). If $M = 1$, then $r_1 = 1$ and

$$A_-(U) = \frac{1}{1+U}, \qquad \tau(U) = \log(1+U).$$

Hence $A_-(U(I)) = e^{-\tau(I)}$, so (C.30) reduces to

$$\frac{p(I)}{1-p(I)} = \frac{p_0}{1-p_0} \frac{e^{-\tau(I)}}{A(I)},$$

which is the one-bad-arm formula.

### C.6 Entropy of the full family distribution

Define the full policy on the active arms by

$$\pi_j^+(I) := (1-p(I))y_j(I), \qquad j=1,\dots,K,$$
$$\pi_m^-(I) := p(I)z_m\big(U(I)\big), \qquad m=1,\dots,M.$$

Its Shannon entropy, denoted $H_{\mathrm{full}} = H_{\mathrm{mode}}$, is

$$H_{\mathrm{full}}(I) := -\sum_{j=1}^{K}\pi_j^+(I)\log\pi_j^+(I) - \sum_{m=1}^{M}\pi_m^-(I)\log\pi_m^-(I). \tag{C.33}$$

We further define the bad-block logarithmic sums

$$C_{-,r}(U) := \sum_{m=1}^{M}\frac{r_m\log r_m}{1+Ur_m}, \qquad C_{-,1}(U) := \sum_{m=1}^{M}\frac{r_m\log(1+Ur_m)}{1+Ur_m},$$
$$D_{-,r}(U) := \sum_{m=1}^{M}\frac{r_m^2\log r_m}{(1+Ur_m)^2}, \qquad D_{-,1}(U) := \sum_{m=1}^{M}\frac{r_m^2\log(1+Ur_m)}{(1+Ur_m)^2}.$$

**Proposition C.11** (Exact full-entropy decomposition)**.** *The bad-block conditional entropy is*

$$H_{\mathrm{bad}}(U) := -\sum_{m=1}^{M}z_m(U)\log z_m(U) = \log A_-(U) + \frac{C_{-,1}(U)-C_{-,r}(U)}{A_-(U)}. \tag{C.34}$$

*Moreover, if*

$$\mathcal{C}_-(U) := \mathrm{Cov}_{m\sim z(U)}\big(z_m,\log z_m\big) = \frac{D_{-,r}(U)-D_{-,1}(U)}{A_-(U)^2} + \frac{B_-(U)}{A_-(U)^3}\big(C_{-,1}(U)-C_{-,r}(U)\big), \tag{C.35}$$

*then*

$$\frac{dH_{\mathrm{bad}}}{dU} = A_-(U)\mathcal{C}_-(U), \qquad \frac{d}{dI}H_{\mathrm{bad}}\big(U(I)\big) = A(I)\mathcal{C}_-\big(U(I)\big). \tag{C.36}$$

*Finally, the full entropy decomposes exactly as*

$$H_{\mathrm{full}}(I) = h_{\mathrm{bin}}\big(p(I)\big) + (1-p(I))H(I) + p(I)H_{\mathrm{bad}}\big(U(I)\big), \tag{C.37}$$

*and its derivative is*

$$\begin{aligned}\frac{dH_{\mathrm{full}}}{dI} = A(I)\Big[&\\ &p(1-p)\big(\mathcal{S}(I)+\mathcal{T}(I)\big)\Big(\mathrm{logit}\,p(I)+H(I)-H_{\mathrm{bad}}\big(U(I)\big)\Big)\\ &-(1-p(I))\mathcal{C}(I)+p(I)\mathcal{C}_-\big(U(I)\big)\Big].\end{aligned} \tag{C.38}$$

*Equivalently,*

$$\begin{aligned}\frac{dH_{\mathrm{full}}}{dp} = &-\mathrm{logit}\,p - H\big(I(p)\big) + H_{\mathrm{bad}}\big(U(I(p))\big)\\ &+\frac{\mathcal{C}\big(I(p)\big)}{p\big(\mathcal{S}(I(p))+\mathcal{T}(I(p))\big)}\\ &-\frac{\mathcal{C}_-\big(U(I(p))\big)}{(1-p)\big(\mathcal{S}(I(p))+\mathcal{T}(I(p))\big)}.\end{aligned} \tag{C.39}$$

*Proof.* The formula (C.34) is the bad-block analogue of (C.27). Indeed, from (C.11),

$$\log z_m(U) = \log r_m - \log(1+Ur_m) - \log A_-(U),$$

and substitution into $-\sum_m z_m\log z_m$ yields (C.34).

Differentiating exactly as in the good-block case gives

$$\frac{dH_{\mathrm{bad}}}{dU} = A_-(U)\mathcal{C}_-(U).$$

Composing with $U(I)$ and using $dU/dI = A(I)/A_-(U(I))$ yields

$$\frac{d}{dI}H_{\mathrm{bad}}\big(U(I)\big) = A(I)\mathcal{C}_-\big(U(I)\big).$$

For the full entropy, expand each block:

$$-\sum_j (1-p)y_j \log\big((1-p)y_j\big) = -(1-p)\log(1-p) + (1-p)H,$$

$$-\sum_m pz_m \log\big(pz_m\big) = -p\log p + pH_{\mathrm{bad}}.$$

Adding the two terms gives

$$H_{\mathrm{full}} = h_{\mathrm{bin}}(p) + (1-p)H + pH_{\mathrm{bad}},$$

which proves (C.37).

Differentiating (C.37),

$$\frac{dH_{\mathrm{full}}}{dI} = -\operatorname{logit} p\,\frac{dp}{dI} - \frac{dp}{dI}H + (1-p)\frac{dH}{dI} + \frac{dp}{dI}H_{\mathrm{bad}} + p\frac{dH_{\mathrm{bad}}}{dI}.$$

Substituting

$$\frac{dp}{dI} = -p(1-p)\big(\mathcal{S}+\mathcal{T}\big)A, \qquad \frac{dH}{dI} = -A\mathcal{C}, \qquad \frac{dH_{\mathrm{bad}}}{dI} = A\mathcal{C}_-,$$

gives (C.38). Dividing by $dp/dI$ yields (C.39). □

*Remark* C.12 (Single bad arm). If $M = 1$, then $z \equiv 1$, $H_{\mathrm{bad}} \equiv 0$, $\mathcal{C}_- \equiv 0$, and $\mathcal{T} \equiv 1$. Thus

$$H_{\mathrm{full}}(I) = h_{\mathrm{bin}}(p(I)) + (1-p(I))H(I),$$

and

$$\frac{dH_{\mathrm{full}}}{dp} = -\operatorname{logit} p - H\big(I(p)\big) + \frac{\mathcal{C}\big(I(p)\big)}{p\big(\mathcal{S}(I(p))+1\big)}.$$

## D Local expansion and asymptotic behavior

Define the good-block power sums and log-power sums by

$$\sigma_m := \sum_{j=1}^{K} q_j^m, \qquad U_m := \sum_{j=1}^{K} q_j^m \log q_j, \tag{D.1}$$

and the bad-block power sums by

$$\rho_m := \sum_{\ell=1}^{M} r_\ell^m. \tag{D.2}$$

Set

$$\Lambda_0 := \sigma_2 + \rho_2.$$

Expanding the good-block scalar sums at $I = 0$ gives

$$\mathcal{C}(0) = \mathrm{Cov}_{j\sim q}(q_j, \log q_j) = U_2 - \sigma_2 U_1 =: C_0 \geq 0, \qquad \mathcal{S}(0) = \sigma_2, \tag{D.3}$$

$$\mathcal{S}(I) = \sigma_2 + K_0 I + O(I^2), \qquad K_0 := 2\sigma_3 - 2\sigma_2^2, \tag{D.4}$$

$$\mathcal{C}(I) = C_0 + C_1 I + O(I^2), \tag{D.5}$$

with

$$C_1 = 2U_3 + \sigma_3 - 3\sigma_2 U_2 - 2\sigma_3 U_1 + 3\sigma_2^2 U_1 - \sigma_2^2. \tag{D.6}$$

The bad block contributes

$$\mathcal{T}(I) = \rho_2 + L_0 I + O(I^2), \qquad L_0 := -2\rho_3 + 2\rho_2^2. \tag{D.7}$$

Indeed, $U(I) = I + O(I^2)$ by (C.16), and a direct expansion of $B_-(U)/A_-(U)^2$ at $U = 0$ gives the stated coefficient.

Introduce the logit coordinate $\theta := \operatorname{logit} p$, for which $\mathrm{d}p/\,\mathrm{d}\theta = p(1-p)$. Then

$$\frac{\mathrm{d}H}{\mathrm{d}\theta} = \frac{\mathrm{d}H}{\mathrm{d}p}\frac{\mathrm{d}p}{\mathrm{d}\theta} = \frac{\mathcal{C}(I)}{\mathcal{S}(I) + \mathcal{T}(I)}. \tag{D.8}$$

**Lemma D.1** (Local value of the entropy drift)**.** *At the initial point* $(p, y, z) = (p_0, q, r)$*,*

$$\left.\frac{\mathrm{d}H}{\mathrm{d}p}\right|_{p=p_0} = \frac{C_0}{p_0(1-p_0)\,\Lambda_0} = \frac{\mathrm{Cov}_{j\sim q}(q_j, \log q_j)}{p_0(1-p_0)\Big(\sum_j q_j^2 + \sum_{m=1}^M r_m^2\Big)}. \tag{D.9}$$

*In particular, if $q$ is uniform on its support (including the vertex case), then $C_0 = 0$; in fact such $q$ are fixed points of* (C.1)*, so $H(\tau)$ is constant for all $\tau$.*

*Proof.* Evaluate (D.8) at $I = 0$ and use (D.3) together with $\mathcal{T}(0) = \rho_2$. Multiplying by $\mathrm{d}\theta/\,\mathrm{d}p = 1/[p(1-p)]$ at $p = p_0$ gives (D.9). □

A first-order Taylor expansion around $p = p_0$ therefore yields

$$\frac{\mathrm{d}H}{\mathrm{d}p} = \frac{C_0}{p_0(1-p_0)\,\Lambda_0} + O(p - p_0), \tag{D.10}$$

where the linear correction can be made explicit from (D.4)–(D.6) and (D.7).

### D.1 Late-time regime as $p \downarrow 0$

Assume $K \geq 2$ and that $q$ has a unique maximizer $q_{j_*} = q_{\max}$. Define

$$\mathcal{A}_\star := \sum_{j \neq j_*} \frac{q_j}{q_{\max} - q_j} > 0, \qquad \varepsilon := 1 - I q_{\max} \downarrow 0. \tag{D.11}$$

The good-block scalar solution gives

$$y_{j_*} = 1 - \mathcal{A}_\star \varepsilon + o(\varepsilon), \qquad y_j = \varepsilon\,\frac{q_j}{q_{\max} - q_j} + o(\varepsilon) \quad (j \neq j_*), \tag{D.12}$$

$$\mathcal{S}(I) = 1 + o(1), \qquad \mathcal{C}(I) = \mathcal{A}_\star \varepsilon\,|\log \varepsilon| + O(\varepsilon). \tag{D.13}$$

The bad block contributes the asymptotics

$$A_-(U) = \frac{M}{U} + O(U^{-2}), \qquad B_-(U) = \frac{M}{U^2} + O(U^{-3}), \qquad \mathcal{T}(I) = \frac{1}{M} + o(1), \tag{D.14}$$

as $U \to \infty$. Using (C.8) and (C.14),

$$e^{\tau(I)} = \frac{1}{\varepsilon}\prod_{j \neq j_*}\Big(1 - \frac{q_j}{q_{\max}}\Big)^{-1}\big(1 + o(1)\big) = \prod_{m=1}^M (1 + U(I) r_m).$$

Since the right-hand side is $U(I)^M \prod_{m=1}^M r_m\,(1 + o(1))$, we obtain

$$U(I) \sim \left[\frac{1}{\varepsilon\,\Big(\prod_{m=1}^M r_m\Big)\Big(\prod_{j \neq j_*}(1 - q_j/q_{\max})\Big)}\right]^{1/M}.$$

Therefore

$$\Xi_M := \frac{p_0}{1-p_0}\,\frac{M}{q_{\max}}\left[\Big(\prod_{m=1}^{M} r_m\Big)\Big(\prod_{j\neq j_*}\Big(1-\frac{q_j}{q_{\max}}\Big)\Big)\right]^{1/M} \in (0,\infty), \tag{D.15}$$

and the exact bad-mass formula (C.30) yields

$$\frac{p}{1-p} = \Xi_M\,\varepsilon^{\,1+\frac{1}{M}}\big(1+O(\varepsilon^{1/M})\big). \tag{D.16}$$

Equivalently, with

$$\alpha := \frac{M}{M+1} \in (0,1), \tag{D.17}$$

one has

$$\varepsilon = \left(\frac{p}{\Xi_M}\right)^{\alpha}\big(1+O(p^{1/(M+1)})\big), \qquad |\log\varepsilon| = \alpha|\log p| + O(1). \tag{D.18}$$

**Lemma D.2** (Late-time asymptotics of the entropy drift)**.** *Under the unique-maximizer assumption with $K \geq 2$, as $p \downarrow 0$,*

$$\frac{\mathrm{d}H}{\mathrm{d}p} \sim \alpha^2\,\frac{\mathcal{A}_\star}{\Xi_M^{\alpha}}\,p^{-\frac{1}{M+1}}\,|\log p|. \tag{D.19}$$

*In particular, the entropy drift diverges like $p^{-1/(M+1)}|\log p|$ near the terminal point.*

*Proof.* Substitute (D.13), (D.14), and (D.18) into (C.32). Since $\mathcal{S}(I)+\mathcal{T}(I) \to 1+\frac{1}{M} = \frac{M+1}{M} = \frac{1}{\alpha}$ and $1-p \to 1$,

$$\frac{\mathrm{d}H}{\mathrm{d}p} \sim \alpha\,\mathcal{A}_\star\,\frac{\varepsilon\,|\log\varepsilon|}{p} \sim \alpha\,\mathcal{A}_\star\frac{1}{p}\left(\frac{p}{\Xi_M}\right)^{\alpha}\alpha|\log p|,$$

which is exactly (D.19). □

*Remark* D.3 (Entropy itself near the terminal vertex). Since

$$\int_0^p u^{\alpha-1}|\log u|\;\mathrm{d}u = \frac{1}{\alpha}p^{\alpha}|\log p| + O(p^{\alpha}),$$

the leading singularity in (D.19) is integrable at 0. More precisely, (D.12) gives $H = \mathcal{A}_\star\varepsilon|\log\varepsilon| + O(\varepsilon)$, and the remainder in (D.18) yields

$$H(p) = \alpha\,\frac{\mathcal{A}_\star}{\Xi_M^{\alpha}}\,p^{\alpha}|\log p| + O(p^{\alpha}). \tag{D.20}$$

If $u$ denotes the uniform distribution on the $K$ good arms, then

$$\mathrm{KL}\big(y(p)\,\|\,u\big) = \log K - \alpha\,\frac{\mathcal{A}_\star}{\Xi_M^{\alpha}}\,p^{\alpha}|\log p| + O(p^{\alpha}). \tag{D.21}$$

More generally, if the full policy is decomposed into bad mass $p$ and good mass $1-p$, then

$$H_{\mathrm{full}}(p) = h_{\mathrm{bin}}(p) + (1-p)H_{\mathrm{good}}(p) + pH_{\mathrm{bad}}(p), \tag{D.22}$$

where $h_{\mathrm{bin}}(p) := -p\log p - (1-p)\log(1-p)$. Since the bad-block dynamics give $H_{\mathrm{bad}}(U(I(p))) \to \log M$ as $p \downarrow 0$, one has

$$H_{\mathrm{full}}(p) = H_{\mathrm{good}}(p) + p|\log p| + O(p),$$

and therefore the leading terminal asymptotics of $H_{\mathrm{full}}$ agree with those of $H_{\mathrm{good}}$ in (D.20).

*Remark* D.4 (A single active good family). If $K = 1$, then $H \equiv 0$ and $\mathcal{A}_\star = 0$. The divergent good-entropy slope and the leading-order equivalence with $H_{\mathrm{full}}$ above do not apply. Instead,

$$H_{\mathrm{full}}(p) = p|\log p| + p(1+\log M) + o(p).$$

# E BATCH AVERAGING AND FAMILY ENTROPY

Promptwise logistic behavior does not automatically imply a logistic batch average. For prompt $i$, write

$$Z_i(t) := \operatorname{logit} p_i(t), \qquad A_i(t) = 1 - p_i(t) = \sigma(-Z_i(t)),$$

where $\sigma(u) = (1 + e^{-u})^{-1}$. For a batch $B$, define

$$\bar{H}_B = \frac{1}{|B|} \sum_{i \in B} H_i, \qquad \bar{A}_B = \frac{1}{|B|} \sum_{i \in B} A_i, \qquad \mu_B = \frac{1}{|B|} \sum_{i \in B} Z_i,$$

and $v_B = |B|^{-1} \sum_{i \in B} (Z_i - \mu_B)^2$. Since $|\sigma'(u)| \le 1/4$, the mean-value theorem and Cauchy–Schwarz give the exact bound

$$\left| \bar{A}_B - \sigma(-\mu_B) \right| \le \frac{1}{4|B|} \sum_{i \in B} |Z_i - \mu_B| \le \frac{\sqrt{v_B}}{4}.$$

Thus, if $|\mu_B - (a_B + b_B \bar{H}_B)| \le \delta$ on a training window, then

$$\left| \bar{A}_B - \sigma(-a_B - b_B \bar{H}_B) \right| \le \frac{\sqrt{v_B} + \delta}{4}.$$

This gives a sufficient condition for an approximate logistic relation after averaging: the mean logit must be nearly affine in mean entropy, and the between-prompt logit variance must be small enough for the bound to be informative.

The full family entropy introduces a separate qualification. For $H_{\text{full}}$, the binary mixing term and the bad-block entropy can make $H_{\text{full}}(t)$ non-monotone. In that case the global curve $(\bar{H}_B(t), \bar{A}_B(t))$ can form a hook rather than a single graph. A logistic approximation can therefore be assessed only on branches where entropy is monotone; it additionally requires an approximately constant logit slope. On monotone branches, the relevant diagnostics are the logit-side slope

$$\Lambda_{\text{full}} = \frac{d}{dH_{\text{full}}} \operatorname{logit} p$$

and the shifted-exponential residual slope

$$\mathcal{B}_{\text{full}} = \frac{d}{dH_{\text{full}}} \log p,$$

since the shifted exponential $R(H) = 1 - ae^H$ is equivalent to $p(H) = ae^H$ in the exact model. The full-entropy simulations illustrate this distinction: windows that cross an entropy turning point can fit poorly to a single curve, whereas monotone windows permit the hazard comparison.

# F PREFIX-TREE LIFT AND TOKEN-LEVEL PROOFS

The prefix-tree identities require a distribution over complete token strings. Let $X$ denote such a string and $F = f(X)$ its family. Because the family is determined by the string,

$$H(X \mid \text{correct}) = H(F \mid \text{correct}) + H(X \mid F, \text{correct}).$$

The scalar solution determines the first term. The within-family term requires additional information and need not vanish when one family dominates. In this appendix only, we impose the stronger specialization that every arm is a distinct complete token string. The resulting prefix identities describe the induced distribution conditional on correctness; the full-policy version appears in Corollary F.18.

## F.1 PREFIX-TREE LIFT OF THE LEAF SOLUTION

Suppose each good arm $j \in [K]$ is a complete token sequence

$$\omega_j = (a_1^{(j)}, \ldots, a_{T_j}^{(j)}, \texttt{eos})$$

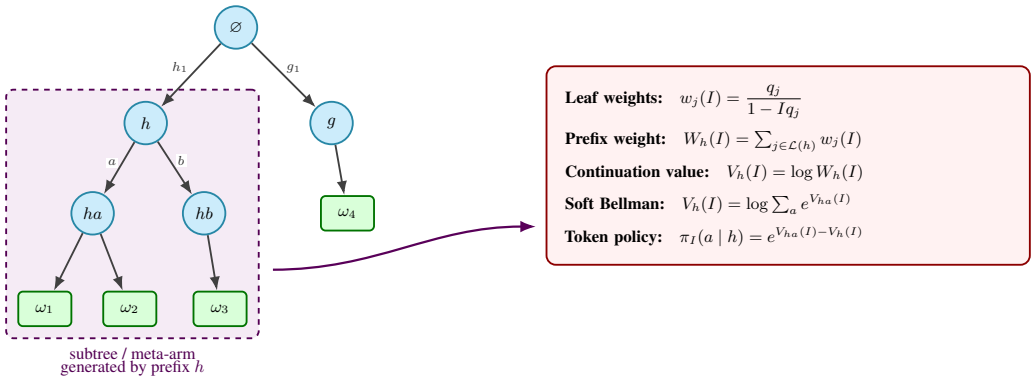


Figure 16: Prefix-tree lift from leaf weights to token policy. A prefix induces a meta-arm whose descendant leaves contribute through the aggregated subtree weight.

over a finite vocabulary, terminated by an explicit end-of-sequence token. Leaves with $q_j = 0$ are inert and may be discarded, so the relevant tree is the finite prefix tree generated by the positive-support sequences. Let $\mathcal{T}$ be that prefix tree, and for a prefix $h$ let

$$\mathcal{L}(h) := \{j \in [K] : h \text{ is a prefix of } \omega_j\}$$

denote the leaf set below $h$. All next-token sums in this appendix run over children present in the active tree, so their subtree weights are positive. Tokens absent from the tree have probability zero and are omitted from logarithms and ratios.

Figure 16 illustrates the construction.

**Lemma F.1** (Prefix masses induced by the leaf solution)**.** *Define the subtree weight*

$$W_h(I) := \sum_{j \in \mathcal{L}(h)} \frac{q_j}{1 - Iq_j}. \tag{F.1}$$

*Then the probability mass of prefix $h$ is exactly*

$$m_h(I) := \sum_{j \in \mathcal{L}(h)} y_j(I) = \frac{W_h(I)}{W_\varnothing(I)} = \frac{W_h(I)}{A(I)}. \tag{F.2}$$

*Proof.* Sum (C.5) over all leaves in $\mathcal{L}(h)$. Since $A(I) = W_\varnothing(I)$, the result is immediate. □

**Proposition F.2** (Induced token-level policy)**.** *For every nonterminal prefix $h$ with $W_h(I) > 0$, the conditional next-token policy satisfies*

$$\pi_I(a \mid h) = \frac{m_{ha}(I)}{m_h(I)} = \frac{W_{ha}(I)}{W_h(I)}. \tag{F.3}$$

*Equivalently, because $I \mapsto p(I)$ is strictly decreasing by Proposition C.9, the same policy can be indexed by the bad mass:*

$$\pi_p(a \mid h) = \pi_{I(p)}(a \mid h) = \frac{W_{ha}(I(p))}{W_h(I(p))}. \tag{F.4}$$

*Proof.* Apply Lemma F.1 to both $h$ and its child $ha$ and use the definition of conditional probability on the prefix tree. □

**Proposition F.3** (Continuation values and a soft Bellman representation)**.** *Define the prefix continuation value*

$$V_h(I) := \log W_h(I). \tag{F.5}$$

*Then for every nonterminal prefix $h$ one has*

$$V_h(I) = \log \sum_a e^{V_{ha}(I)}, \qquad \pi_I(a \mid h) = \exp\big(V_{ha}(I) - V_h(I)\big). \tag{F.6}$$

*Moreover, with*

$$\lambda_h(I) := \frac{W_h'(I)}{W_h(I)} = \frac{\sum_{j\in\mathcal{L}(h)} \dfrac{q_j^2}{(1-Iq_j)^2}}{\sum_{j\in\mathcal{L}(h)} \dfrac{q_j}{1-Iq_j}}, \tag{F.7}$$

*one has the value-slope identity*

$$\frac{\mathrm{d}}{\mathrm{d}I} V_h(I) = \lambda_h(I) = \sum_a \pi_I(a \mid h)\lambda_{ha}(I). \tag{F.8}$$

*Proof.* Since $W_h(I) = \sum_a W_{ha}(I)$, taking logarithms gives the soft Bellman relation in (F.6); the policy formula is exactly Proposition F.2 rewritten in terms of $V_h$. Differentiating $V_h = \log W_h$ gives $\frac{\mathrm{d}}{\mathrm{d}I} V_h = W_h'/W_h = \lambda_h$. Finally,

$$\sum_a \pi_I(a \mid h)\lambda_{ha}(I) = \sum_a \frac{W_{ha}(I)}{W_h(I)} \frac{W_{ha}'(I)}{W_{ha}(I)} = \frac{\sum_a W_{ha}'(I)}{W_h(I)} = \frac{W_h'(I)}{W_h(I)} = \lambda_h(I). \qquad \square$$

### F.2 Exact token flows and the hierarchy of entropy drift

**Proposition F.4** (Entropy chain rule on the prefix tree)**.** *For each nonterminal prefix $h$, define the local token entropy*

$$H_h(I) := -\sum_a \pi_I(a \mid h) \log \pi_I(a \mid h). \tag{F.9}$$

*Then the sequence-level entropy studied above admits the exact decomposition*

$$H(I) = -\sum_{j=1}^{K} y_j(I) \log y_j(I) = \sum_{h\in\mathcal{T}_{\mathrm{int}}} m_h(I)\, H_h(I), \tag{F.10}$$

*where $\mathcal{T}_{\mathrm{int}}$ is the set of nonterminal prefixes. Equivalently,*

$$H(I) = \sum_{t\geq 1} \sum_{|h|=t-1} m_h(I) H_h(I) = \sum_{t\geq 1} \rho_t(I)\, \mathbb{E}_{h\sim \bar{d}_t^{\pi_I}} \big[H_h(I)\big], \tag{F.11}$$

*where*

$$\rho_t(I) := \sum_{|h|=t-1} m_h(I)$$

*is the probability of reaching depth $t$, and*

$$\bar{d}_t^{\pi_I}(h) := \frac{m_h(I)}{\rho_t(I)}$$

*is the normalized depth-$t$ occupancy distribution whenever $\rho_t(I) > 0$.*

*Proof.* Write $h_t^{(j)} := (a_1^{(j)}, \ldots, a_t^{(j)})$ with $h_0^{(j)} := \varnothing$. By repeated conditioning,

$$y_j(I) = \prod_{t=1}^{T_j+1} \pi_I\big(a_t^{(j)} \mid h_{t-1}^{(j)}\big),$$

where the final token is `eos`. Taking logarithms and summing over leaves gives

$$H(I) = -\sum_j y_j(I) \sum_{t=1}^{T_j+1} \log \pi_I\big(a_t^{(j)} \mid h_{t-1}^{(j)}\big).$$

Regrouping terms by prefix $h$ and next token $a$ yields

$$H(I) = -\sum_h \sum_a \Big( \sum_{j\in\mathcal{L}(ha)} y_j(I) \Big) \log \pi_I(a \mid h) = -\sum_h m_h(I) \sum_a \pi_I(a \mid h) \log \pi_I(a \mid h),$$

which is (F.10). Grouping the same sum by depth gives (F.11). $\square$

*Remark* F.5. Equation (F.10) shows that the sequence-level entropy is exactly the occupancy-weighted sum of token-level conditional entropies. The scalar curve $H(p)$ alone does not determine token-level dynamics; the bridge is available here because the explicit leaf distribution (C.5) is known.

**Lemma F.6** (Exact prefix and token flows)**.** *For every nonterminal prefix $h$, the induced prefix masses and token conditionals obey*

$$\frac{\mathrm{d}}{\mathrm{d}I} m_h(I) = m_h(I)\big(\lambda_h(I) - \lambda_\varnothing(I)\big), \tag{F.12}$$

$$\frac{\mathrm{d}}{\mathrm{d}I} \pi_I(a \mid h) = \pi_I(a \mid h)\big(\lambda_{ha}(I) - \lambda_h(I)\big), \tag{F.13}$$

*and the local entropy satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}I} H_h(I) = -\operatorname{Cov}_{a\sim\pi_I(\cdot|h)}\big(\lambda_{ha}(I), \log \pi_I(a \mid h)\big). \tag{F.14}$$

*In the $p$-parameterization,*

$$\frac{\mathrm{d}}{\mathrm{d}p} \pi_p(a \mid h) = -\frac{\pi_p(a \mid h)\big(\lambda_{ha}(I(p)) - \lambda_h(I(p))\big)}{p(1-p)\big(\mathcal{S}(I(p)) + \mathcal{T}(I(p))\big)A(I(p))}. \tag{F.15}$$

*Proof.* Differentiate (F.2):

$$\frac{\mathrm{d}}{\mathrm{d}I} \log m_h(I) = \frac{W_h'(I)}{W_h(I)} - \frac{A'(I)}{A(I)} = \lambda_h(I) - \lambda_\varnothing(I),$$

which gives (F.12). Likewise,

$$\frac{\mathrm{d}}{\mathrm{d}I} \log \pi_I(a \mid h) = \frac{W_{ha}'(I)}{W_{ha}(I)} - \frac{W_h'(I)}{W_h(I)} = \lambda_{ha}(I) - \lambda_h(I),$$

and multiplying by $\pi_I(a \mid h)$ gives (F.13). Since

$$\sum_a \pi_I(a \mid h)\lambda_{ha}(I) = \lambda_h(I)$$

by Proposition F.3, (F.14) follows from

$$\frac{\mathrm{d}}{\mathrm{d}I} H_h(I) = -\sum_a \frac{\mathrm{d}}{\mathrm{d}I} \pi_I(a \mid h) \log \pi_I(a \mid h).$$

Finally, (C.31) implies

$$\frac{\mathrm{d}p}{\mathrm{d}I} = p(1-p)\frac{\mathrm{d}}{\mathrm{d}I} \operatorname{logit} p(I) = -p(1-p)\big(\mathcal{S}(I) + \mathcal{T}(I)\big)A(I),$$

which yields (F.15) by the chain rule. □

**Proposition F.7** (Hierarchical decomposition of the entropy drift)**.** *The macroscopic sequence-level covariance from* (C.26) *admits the exact prefix-tree decomposition*

$$\mathcal{C}(I) = \frac{1}{A(I)} \sum_{h\in\mathcal{T}_{\mathrm{int}}} m_h(I)\Big[\operatorname{Cov}_{a\sim\pi_I(\cdot|h)}\big(\lambda_{ha}(I), \log \pi_I(a \mid h)\big) - \big(\lambda_h(I) - \lambda_\varnothing(I)\big)H_h(I)\Big]. \tag{F.16}$$

*Proof.* Proposition F.4 gives

$$\frac{\mathrm{d}H}{\mathrm{d}I} = \sum_{h\in\mathcal{T}_{\mathrm{int}}} m_h(I)\frac{\mathrm{d}H_h}{\mathrm{d}I} + \sum_{h\in\mathcal{T}_{\mathrm{int}}} \frac{\mathrm{d}m_h}{\mathrm{d}I} H_h(I).$$

Substitute (F.12) and (F.14) from Lemma F.6. On the other hand, Proposition C.9 gives

$$\frac{\mathrm{d}H}{\mathrm{d}I} = \frac{\mathrm{d}H/\,\mathrm{d}\tau}{\mathrm{d}I/\,\mathrm{d}\tau} = -A(I)\mathcal{C}(I).$$

Equating the two formulas and dividing by $-A(I)$ yields (F.16). □

*Remark* F.8 (Local policy change and prefix occupancy). Equation (F.16) separates changes in local token entropy from changes in prefix occupancy. Neither contribution has a fixed sign at an individual prefix: a local token policy can become more diffuse even while the total sequence entropy decreases. Their occupancy-weighted sum equals the nonnegative good-block covariance.

**Proposition F.9** (Closed form of the local token entropy)**.** *For every nonterminal prefix $h$, the local token entropy admits the exact representation*

$$H_h(I) = -\sum_a \frac{W_{ha}(I)}{W_h(I)} \log \frac{W_{ha}(I)}{W_h(I)} = \log W_h(I) - \frac{1}{W_h(I)} \sum_a W_{ha}(I) \log W_{ha}(I). \tag{F.17}$$

*Moreover, on every interval where $\mathcal{C}(I) > 0$, the global entropy $H(I)$ may be used as a scalar clock, and one has the exact differential relation*

$$\frac{\mathrm{d}H_h}{\mathrm{d}H} = \frac{\mathrm{Cov}_{a\sim\pi_I(\cdot|h)}\big(\lambda_{ha}(I), \log \pi_I(a \mid h)\big)}{A(I)\mathcal{C}(I)}. \tag{F.18}$$

*Equivalently,*

$$\frac{\mathrm{d}H_h}{\mathrm{d}p} = \frac{\mathrm{Cov}_{a\sim\pi_I(\cdot|h)}\big(\lambda_{ha}(I), \log \pi_I(a \mid h)\big)}{p(1-p)\big(\mathcal{S}(I) + \mathcal{T}(I)\big)A(I)}. \tag{F.19}$$

*Proof.* The first identity is just (F.9) together with (F.3). The formulas (F.18) and (F.19) follow by combining (F.14) with (C.28) and with $\frac{\mathrm{d}p}{\mathrm{d}I} = -p(1-p)(\mathcal{S}(I) + \mathcal{T}(I))A(I)$. □

**Corollary F.10** (Occupancy-averaged token entropy)**.** *Define the expected number of token decisions*

$$L(I) := \sum_{h\in\mathcal{T}_{\mathrm{int}}} m_h(I) = \sum_{j=1}^{K} y_j(I)\,\ell_j, \qquad \ell_j := T_j + 1, \tag{F.20}$$

*and the occupancy-averaged token entropy*

$$\bar{H}_{\mathrm{tok}}(I) := \frac{1}{L(I)} \sum_{h\in\mathcal{T}_{\mathrm{int}}} m_h(I) H_h(I). \tag{F.21}$$

*Then*

$$\bar{H}_{\mathrm{tok}}(I) = \frac{H(I)}{L(I)}. \tag{F.22}$$

*Hence*

$$\frac{\mathrm{d}\bar{H}_{\mathrm{tok}}}{\mathrm{d}p} = \frac{1}{L(I(p))}\frac{\mathrm{d}H}{\mathrm{d}p} - \frac{H(I(p))}{L(I(p))^2}\frac{\mathrm{d}L}{\mathrm{d}p}, \tag{F.23}$$

*where*

$$\frac{\mathrm{d}L}{\mathrm{d}p} = -\frac{\mathrm{Cov}_{j\sim y(I(p))}\Big(\dfrac{q_j}{1-I(p)q_j},\, \ell_j\Big)}{p(1-p)\big(\mathcal{S}(I(p)) + \mathcal{T}(I(p))\big)A(I(p))}. \tag{F.24}$$

*If all sequences contain $T$ token decisions including termination, then $L(I) \equiv T$ and therefore*

$$\bar{H}_{\mathrm{tok}}(p) = \frac{H(p)}{T}, \qquad \frac{\mathrm{d}\bar{H}_{\mathrm{tok}}}{\mathrm{d}p} = \frac{1}{T}\frac{\mathrm{d}H}{\mathrm{d}p}. \tag{F.25}$$

*Proof.* The identity (F.22) is immediate from (F.10). Also,

$$L(I) = \sum_j y_j(I)\ell_j = \frac{\sum_j \ell_j\, q_j/(1-Iq_j)}{A(I)}.$$

Differentiating $L(I) = \sum_j y_j(I)\ell_j$ gives

$$\frac{\mathrm{d}L}{\mathrm{d}I} = \mathrm{Cov}_{j\sim y(I)}\Big(\frac{q_j}{1-Iq_j}, \ell_j\Big),$$

and (F.24) follows by dividing by $\mathrm{d}p/\,\mathrm{d}I$. □

*Remark* F.11 (Dependence on the entropy estimator). The identity $H = \sum_h m_h H_h$ does not determine every local entropy from the total. The ratio $H/L$ applies to an average that pools token decisions with their on-policy occupancies, including the specified termination convention. An average that first normalizes each sequence by its own length is generally different. Comparisons with logged entropy must also match the conditioning on correctness and the string-level policy. If $H$ decreases while $L$ increases, $H/L$ decreases; a rise in this ratio requires length to fall sufficiently quickly or a change in a component not represented by this conditional leaf model.

### F.3 Asymptotic geometry: collapse and freezing

**Proposition F.12** (Token-level collapse along the winning path)**.** *Assume $K \geq 2$ and that $q$ has a unique maximizer $q_{j_*} = q_{\max}$ and let*

$$\omega_* = (a_1^*, \ldots, a_T^*, \texttt{eos})$$

*be the corresponding winning leaf sequence. For $t \in \{1, \ldots, T+1\}$, let*

$$h_t^* := (a_1^*, \ldots, a_{t-1}^*)$$

*be the on-path prefix, and define the branch constant*

$$\Gamma_t := \sum_{j \in \mathcal{L}(h_t^*) \setminus \mathcal{L}(h_t^* a_t^*)} \frac{q_j}{q_{\max} - q_j}, \tag{F.26}$$

*with the convention that $a_{T+1}^* = \texttt{eos}$. Then, as $I \uparrow 1/q_{\max}$ (equivalently $\varepsilon \downarrow 0$), the following hold:*

1. *for every prefix $h$ that is not on the winning path, $m_h(I) = O(\varepsilon)$;*
2. *along the winning path,*

$$1 - \pi_I(a_t^* \mid h_t^*) = \Gamma_t \varepsilon + o(\varepsilon), \qquad H_{h_t^*}(I) = \Gamma_t \varepsilon \, |\log \varepsilon| + O(\varepsilon); \tag{F.27}$$

3. *in the $p$-parameterization,*

$$1 - \pi_p(a_t^* \mid h_t^*) = \Gamma_t \, \Xi_M^{-\alpha} p^\alpha + o(p^\alpha), \qquad H_{h_t^*}(p) = \alpha \, \Gamma_t \, \Xi_M^{-\alpha} p^\alpha \, |\log p| + O(p^\alpha). \tag{F.28}$$

*In particular, the token policy becomes deterministic along the winning path, while off-path prefixes lose all occupancy.*

*Proof.* Write $w_j(I) := q_j/(1 - Iq_j)$. For the winning leaf, $w_{j_*}(I) = q_{\max}/\varepsilon$. For every $j \neq j_*$,

$$w_j(I) = \frac{q_j}{1 - Iq_j} = \frac{q_{\max} q_j}{q_{\max} - q_j} + O(\varepsilon).$$

Hence if $h$ is not a prefix of $\omega_*$, then $j_* \notin \mathcal{L}(h)$ and therefore $W_h(I) = O(1)$, whereas

$$A(I) = W_\varnothing(I) = \frac{q_{\max}}{\varepsilon} + O(1),$$

so $m_h(I) = W_h(I)/A(I) = O(\varepsilon)$. This proves the first claim.

Now fix an on-path prefix $h_t^*$. The child subtree corresponding to the winning next token satisfies

$$W_{h_t^* a_t^*}(I) = \frac{q_{\max}}{\varepsilon} + O(1),$$

whereas the total weight of the competing children is

$$\sum_{a \neq a_t^*} W_{h_t^* a}(I) = q_{\max} \Gamma_t + o(1)$$

by the definition of $\Gamma_t$. Since $W_{h_t^*}(I) = q_{\max}/\varepsilon + O(1)$, we obtain

$$1 - \pi_I(a_t^* \mid h_t^*) = \frac{\sum_{a \neq a_t^*} W_{h_t^* a}(I)}{W_{h_t^*}(I)} = \Gamma_t \varepsilon + o(\varepsilon).$$

Let $\delta_t(I) := 1 - \pi_I(a_t^* \mid h_t^*)$. The local distribution at $h_t^*$ has the form $(1-\delta_t, \delta_t \nu_t)$ for some probability vector $\nu_t$ supported on the losing children. Therefore,

$$H_{h_t^*}(I) = -(1-\delta_t)\log(1-\delta_t) - \delta_t \log \delta_t + \delta_t H(\nu_t) = \delta_t |\log \delta_t| + O(\delta_t),$$

which gives the second claim when $\Gamma_t > 0$. If $\Gamma_t = 0$, the prefix has only the winning child on the active tree, so its local entropy is identically zero.

Finally, substitute (D.18) into the previous asymptotics and use $|\log \varepsilon| = \alpha |\log p| + O(1)$ to obtain (F.28). □

**Proposition F.13** (Freezing of off-path conditionals)**.** *Assume that $q$ has a unique maximizer $q_{j_*} = q_{\max}$ and let $g$ be a prefix that is not on the winning path, so that $j_* \notin \mathcal{L}(g)$. Define the conditional distribution over leaves in the subtree of $g$ by*

$$\nu_{I,g}(j) := \Pr(\omega_j \mid g) = \frac{y_j(I)}{m_g(I)} = \frac{w_j(I)}{W_g(I)}, \qquad j \in \mathcal{L}(g), \tag{F.29}$$

*where $w_j(I) = q_j/(1 - Iq_j)$. Then as $I \uparrow 1/q_{\max}$ one has*

$$\nu_{I,g}(j) \longrightarrow \bar{\nu}_g(j) := \frac{\dfrac{q_j}{q_{\max} - q_j}}{\displaystyle\sum_{\ell \in \mathcal{L}(g)} \frac{q_\ell}{q_{\max} - q_\ell}}, \qquad j \in \mathcal{L}(g) \text{ with } q_j > 0. \tag{F.30}$$

*In particular, if $q_j > 0$ for all $j \in \mathcal{L}(g)$, then $\bar{\nu}_g$ is strictly positive on the whole subtree. The induced token policy at $g$ therefore converges to the frozen limit*

$$\pi_I(a \mid g) \longrightarrow \bar{\pi}_g(a) := \frac{\displaystyle\sum_{j \in \mathcal{L}(ga)} \frac{q_j}{q_{\max} - q_j}}{\displaystyle\sum_{\ell \in \mathcal{L}(g)} \frac{q_\ell}{q_{\max} - q_\ell}}. \tag{F.31}$$

*Consequently, $H_g(I)$ converges to a finite limit*

$$H_g^{\mathrm{fr}} := -\sum_a \bar{\pi}_g(a) \log \bar{\pi}_g(a) \geq 0, \tag{F.32}$$

*which is strictly positive whenever at least two children of $g$ carry positive frozen mass. Moreover,*

$$m_g(I) H_g(I) \longrightarrow 0. \tag{F.33}$$

*Proof.* Because $j_* \notin \mathcal{L}(g)$, every denominator $1 - Iq_j$ with $j \in \mathcal{L}(g)$ stays bounded away from zero as $I \uparrow 1/q_{\max}$. Hence

$$w_j(I) = \frac{q_j}{1 - Iq_j} \longrightarrow q_{\max} \frac{q_j}{q_{\max} - q_j} \qquad (j \in \mathcal{L}(g)),$$

and summing over $j \in \mathcal{L}(g)$ gives

$$W_g(I) \longrightarrow q_{\max} \sum_{\ell \in \mathcal{L}(g)} \frac{q_\ell}{q_{\max} - q_\ell}.$$

Taking the ratio $w_j(I)/W_g(I)$ proves (F.30); summing those limits over children gives (F.31). The local entropy is a continuous function of the finite-dimensional token simplex, so (F.32) follows. Finally, Proposition F.12 gives $m_g(I) = O(\varepsilon)$ for every off-path prefix, while $0 \leq H_g(I) \leq \log |\mathrm{ch}(g)|$ is bounded, which yields (F.33). □

*Remark* F.14 (Vanishing occupancy and residual conditional entropy). Propositions F.12 and F.13 together complete the asymptotic picture. Along the winning path the token policy becomes deterministic, whereas inside extinct branches the conditional policy can retain nonzero residual entropy and freeze to a finite distribution. Sequence-level entropy vanishes because the occupancy of those branches goes to zero, not because every local policy becomes deterministic.

**Proposition F.15** (Late-time local entropy collapse on the winning path)**.** *Assume the unique-maximizer setting with $K \geq 2$ from Appendix D, and let $j_*$ be the unique leaf with $q_{j_*} = q_{\max}$. For a nonterminal prefix $h \prec \omega_{j_*}$ on the winning path, let $a_h^*$ be the next token on $\omega_{j_*}$ and define*

$$\Gamma_h := \sum_{a \neq a_h^*} \sum_{j \in \mathcal{L}(ha)} \frac{q_j}{q_{\max} - q_j}. \tag{F.34}$$

*Then, with $\alpha = \frac{M}{M+1}$ and $\Xi_M$ from* (D.15)*,*

$$H_h(p) = \alpha \, \frac{\Gamma_h}{\Xi_M^\alpha} \, p^\alpha |\log p| + O(p^\alpha) \qquad (p \downarrow 0). \tag{F.35}$$

*Consequently,*

$$\frac{\mathrm{d}H_h}{\mathrm{d}p} = \alpha^2 \, \frac{\Gamma_h}{\Xi_M^\alpha} \, p^{-\frac{1}{M+1}} |\log p| + O\Big(p^{-\frac{1}{M+1}}\Big), \tag{F.36}$$

*so each winning-path prefix with $\Gamma_h > 0$ has the same entropy exponent as $H(p)$. A prefix with $\Gamma_h = 0$ has zero local entropy throughout.*

*Proof.* Let $\varepsilon = 1 - I q_{\max}$. For a child $a \neq a_h^*$, define

$$\Gamma_{h,a} := \sum_{j \in \mathcal{L}(ha)} \frac{q_j}{q_{\max} - q_j}, \qquad \Gamma_h = \sum_{a \neq a_h^*} \Gamma_{h,a}.$$

The same expansion used in Proposition F.12 gives

$$\pi_I(a_h^* \mid h) = 1 - \Gamma_h \varepsilon + O(\varepsilon^2), \qquad \pi_I(a \mid h) = \Gamma_{h,a} \varepsilon + O(\varepsilon^2) \quad (a \neq a_h^*),$$

where terms with $\Gamma_{h,a} = 0$ may be omitted. Substituting this finite-dimensional expansion into the exact local entropy definition gives

$$H_h(I) = \Gamma_h \varepsilon |\log \varepsilon| + O(\varepsilon).$$

Using (D.18), namely $\varepsilon \sim (p/\Xi_M)^\alpha$ and $|\log \varepsilon| = \alpha |\log p| + O(1)$, yields (F.35).

For the derivative, we avoid differentiating the asymptotic remainder. Instead use the exact covariance identity (F.14). Along the winning child,

$$\lambda_{h a_h^*}(I) = \frac{q_{\max}}{\varepsilon} + O(1),$$

whereas every losing child has $\lambda_{ha}(I) = O(1)$. Combining these rate expansions with the displayed expansion of $\pi_I(\cdot \mid h)$ in (F.14) gives

$$\frac{\mathrm{d}H_h}{\mathrm{d}I} = -q_{\max} \Gamma_h |\log \varepsilon| + O(1).$$

Using the exact identity $\mathrm{d}p/\,\mathrm{d}I = -p(1-p)(\mathcal{S} + \mathcal{T})A$ and the scalar expansions,

$$\frac{\mathrm{d}p}{\mathrm{d}I} = -\frac{q_{\max} \Xi_M}{\alpha} \varepsilon^{1/\alpha - 1} \big(1 + O(\varepsilon^{1/M})\big).$$

Dividing the two exact derivatives and rewriting in the $p$-coordinate gives

$$\frac{\mathrm{d}H_h}{\mathrm{d}p} = \alpha \Gamma_h \Xi_M^{-1} \varepsilon^{1 - 1/\alpha} |\log \varepsilon| + O\Big(\varepsilon^{1 - 1/\alpha}\Big) = \alpha^2 \frac{\Gamma_h}{\Xi_M^\alpha} p^{\alpha - 1} |\log p| + O(p^{\alpha - 1}),$$

which is (F.36) because $\alpha - 1 = -1/(M+1)$. □

*Remark* F.16. The constants $\Gamma_h$ split the global late-time coefficient across the successive branch points of the winning path. In particular,

$$\sum_{h \prec \omega_{j_*}} \Gamma_h = \mathcal{A}_\star,$$

so the leading-order sequence entropy is the sum of the leading local entropies along the winning path. Off-path prefixes may retain nonzero conditional entropy, but their occupancy $m_h(p)$ vanishes, so they do not contribute at leading order to the on-policy token entropy observed in rollouts.

**Proposition F.17** (Terminal equivalence of full and good-block entropy)**.** *Assume $K \geq 2$ and that $q$ has a unique maximizer $q_{j_*} = q_{\max}$, and let $\alpha$, $\Xi_M$, and $\mathcal{A}_\star$ be defined by* (D.17)*,* (D.15)*, and* (D.11)*. Set*

$$\bar{s} := \frac{1}{M}\sum_{m=1}^{M}\frac{1}{r_m}, \qquad \Sigma_-^2 := \sum_{m=1}^{M}\Big(\frac{1}{r_m} - \bar{s}\Big)^2.$$

*Then, as $U \to \infty$,*

$$z_m(U) = \frac{1}{M} + \frac{\bar{s} - 1/r_m}{MU} + O(U^{-2}), \tag{F.37}$$

$$H_{\mathrm{bad}}(U) = \log M - \frac{\Sigma_-^2}{2M}\,U^{-2} + O(U^{-3}), \qquad \mathcal{C}_-(U) = \frac{\Sigma_-^2}{M^2}\,U^{-2} + O(U^{-3}). \tag{F.38}$$

*Consequently, using $U(I(p)) \asymp p^{-1/(M+1)}$, one has*

$$H_{\mathrm{bad}}\big(U(I(p))\big) = \log M + O\big(p^{2/(M+1)}\big), \qquad \mathcal{C}_-\big(U(I(p))\big) = O\big(p^{2/(M+1)}\big). \tag{F.39}$$

*Moreover,*

$$H_{\mathrm{full}}(p) = H(p) + p|\log p| + p\big(1 + \log M\big) + o(p), \tag{F.40}$$

*and therefore*

$$H_{\mathrm{full}}(p) \sim H(p) \sim \alpha\,\frac{\mathcal{A}_\star}{\Xi_M^\alpha}\,p^\alpha|\log p|. \tag{F.41}$$

*Finally,*

$$\frac{dH_{\mathrm{full}}}{dp} = \frac{dH}{dp} + |\log p| + \log M + o(1), \tag{F.42}$$

*so in particular*

$$\frac{dH_{\mathrm{full}}}{dp} \sim \frac{dH}{dp} \sim \alpha^2\,\frac{\mathcal{A}_\star}{\Xi_M^\alpha}\,p^{-1/(M+1)}|\log p|. \tag{F.43}$$

*Proof.* From

$$\frac{r_m}{1 + U r_m} = \frac{1}{U} - \frac{1}{r_m U^2} + O(U^{-3})$$

one gets

$$A_-(U) = \sum_{m=1}^{M}\frac{r_m}{1 + U r_m} = \frac{M}{U} - \frac{1}{U^2}\sum_{m=1}^{M}\frac{1}{r_m} + O(U^{-3}),$$

which implies (F.37) after normalization. Since the entropy is smooth and strictly concave at the uniform distribution,

$$H(u + \delta) = \log M - \frac{M}{2}\sum_{m=1}^{M}\delta_m^2 + O(\|\delta\|_3^3) \qquad (u_m \equiv 1/M,\ \sum_m \delta_m = 0),$$

substituting

$$\delta_m = \frac{\bar{s} - 1/r_m}{MU} + O(U^{-2})$$

gives the first expansion in (F.38). Expanding the covariance directly at $z_m = 1/M + \delta_m$ gives $\mathcal{C}_- = \sum_m \delta_m^2 + O(\|\delta\|^3)$, which proves the second expansion without differentiating a remainder.

Because $U(I(p)) \asymp p^{-1/(M+1)}$, (F.39) is immediate. Next, using the exact decomposition

$$H_{\mathrm{full}}(p) = h_{\mathrm{bin}}(p) + (1 - p)H(p) + pH_{\mathrm{bad}}\big(U(I(p))\big)$$

and the standard expansion

$$h_{\mathrm{bin}}(p) = p|\log p| + p + O(p^2),$$

together with $H(p) = o(1)$ and (F.39), yields

$$H_{\mathrm{full}}(p) - H(p) = p|\log p| + p\big(1 + \log M\big) + o(p),$$

which is (F.40). Since $\alpha \in (0, 1)$, one has $p|\log p| = o\big(p^\alpha|\log p|\big)$ and $p = o\big(p^\alpha|\log p|\big)$, so combining with (D.20) proves (F.41).

For the derivative, start from (C.39):

$$\frac{dH_{\mathrm{full}}}{dp} = -\operatorname{logit} p - H(p) + H_{\mathrm{bad}}\big(U(I(p))\big) + \frac{\mathcal{C}(I(p))}{p(\mathcal{S}(I(p)) + \mathcal{T}(I(p)))} - \frac{\mathcal{C}_-(U(I(p)))}{(1-p)(\mathcal{S}(I(p)) + \mathcal{T}(I(p)))}.$$

Using

$$\frac{dH}{dp} = \frac{\mathcal{C}(I(p))}{p(1-p)(\mathcal{S}(I(p)) + \mathcal{T}(I(p)))}$$

we rewrite the covariance term as

$$\frac{\mathcal{C}(I(p))}{p(\mathcal{S}(I(p)) + \mathcal{T}(I(p)))} = (1-p)\frac{dH}{dp} = \frac{dH}{dp} - p\frac{dH}{dp}.$$

By Lemma D.2, one has

$$p\frac{dH}{dp} = O\big(p^{\alpha}|\log p|\big) = o(1),$$

so

$$\frac{dH_{\mathrm{full}}}{dp} = \frac{dH}{dp} - \operatorname{logit} p - H(p) + H_{\mathrm{bad}}\big(U(I(p))\big) + o(1).$$

Now $-\operatorname{logit} p = |\log p| + o(1)$, $H(p) \to 0$, and $H_{\mathrm{bad}}(U(I(p))) = \log M + o(1)$, giving (F.42). Since $\frac{dH}{dp} \sim \alpha^2 \frac{\mathcal{A}_\star}{\Xi_M^\alpha} p^{-1/(M+1)}|\log p|$ from Lemma D.2, the additive correction in (F.42) is lower order, proving (F.43). □

**Corollary F.18** (Full token-entropy bridge)**.** *Under the assumptions of Proposition F.17, suppose the active good and bad arms are distinct complete token sequences on a common finite prefix tree, and let $m_h^{\mathrm{full}}(I)$ and $H_h^{\mathrm{full}}(I)$ denote the induced prefix occupancies and local token entropies under the full policy. Then*

$$H_{\mathrm{full}}(I) = \sum_{h \in \mathcal{T}_{\mathrm{int}}^{\mathrm{full}}} m_h^{\mathrm{full}}(I)\, H_h^{\mathrm{full}}(I). \tag{F.44}$$

*If*

$$L_{\mathrm{full}}(I) := \sum_{h \in \mathcal{T}_{\mathrm{int}}^{\mathrm{full}}} m_h^{\mathrm{full}}(I)$$

*is the expected number of token decisions and $L_{\mathrm{full}}(p) \to L_\star \in (0,\infty)$ as $p \downarrow 0$, then the occupancy-averaged full token entropy*

$$\bar{H}_{\mathrm{tok}}^{\mathrm{full}}(p) := \frac{H_{\mathrm{full}}(p)}{L_{\mathrm{full}}(p)}$$

*satisfies*

$$\bar{H}_{\mathrm{tok}}^{\mathrm{full}}(p) \sim \frac{1}{L_\star}\alpha\, \frac{\mathcal{A}_\star}{\Xi_M^\alpha}\, p^{\alpha}|\log p|. \tag{F.45}$$

*If in addition $\frac{dL_{\mathrm{full}}}{dp} = O\big(p^{-1/(M+1)}\big)$, then*

$$\frac{d\bar{H}_{\mathrm{tok}}^{\mathrm{full}}}{dp} \sim \frac{1}{L_\star}\alpha^2\, \frac{\mathcal{A}_\star}{\Xi_M^\alpha}\, p^{-1/(M+1)}|\log p|. \tag{F.46}$$

# G Hazard identities and exponential approximations

## G.1 Exact hazard identities

We use *hazard* for the derivative of a transformed bad mass with respect to entropy. Throughout this section, assume the good block is nonuniform, so $\mathcal{C}(I) > 0$ at every finite time and $H$ is a valid

trajectory coordinate. For a uniform good block, $H$ is constant and these derivatives with respect to $H$ are undefined. In internal time, the entropy drift satisfies

$$\frac{\mathrm{d}H}{\mathrm{d}\tau} = -\mathcal{C}(I(\tau)), \tag{G.1}$$

because (C.28) and (C.6) give $\mathrm{d}H/\,\mathrm{d}\tau = (\,\mathrm{d}H/\,\mathrm{d}I)(\,\mathrm{d}I/\,\mathrm{d}\tau) = -\mathcal{C}$. Meanwhile the bad mass obeys

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathrm{logit}\,p = -(s_2 + t_2) = -(\mathcal{S}(I(\tau)) + \mathcal{T}(I(\tau))). \tag{G.2}$$

Dividing (G.2) by (G.1) yields the exact logit identity

$$\frac{\mathrm{d}}{\mathrm{d}H}\,\mathrm{logit}\,p = \frac{\mathcal{S}(I) + \mathcal{T}(I)}{\mathcal{C}(I)} =: \Lambda(H). \tag{G.3}$$

The shifted exponential bridge $R(H) = 1 - a\exp(H)$ is equivalent, in the exact model, to $p(H) = a\exp(H)$. Its residual hazard is therefore

$$\frac{\mathrm{d}}{\mathrm{d}H}\,\log p = \frac{(1-p)(\mathcal{S}(I) + \mathcal{T}(I))}{\mathcal{C}(I)} =: \mathcal{B}(H). \tag{G.4}$$

For comparison, the unshifted accuracy hazard is

$$-\frac{\mathrm{d}}{\mathrm{d}H}\,\log(1-p) = \frac{p(\mathcal{S}(I) + \mathcal{T}(I))}{\mathcal{C}(I)}. \tag{G.5}$$

An approximately constant residual hazard $\mathcal{B}(H) \approx b$ gives the free-slope relation $p(H) \approx ae^{bH}$. The unit-slope relation above additionally requires $b \approx 1$ in the chosen entropy units. An approximately constant logit hazard gives an affine logit and hence a sigmoid relation.

**Corollary G.1** (Shifted-exponential and logistic envelopes on finite windows)**.** *Fix an entropy interval $[H_1, H_2]$ on a nonuniform good-block trajectory. If*

$$\epsilon_- \le \mathcal{B}(H) \le \epsilon_+ \qquad \textit{for all } H \in [H_1, H_2],$$

*then for all $H \in [H_1, H_2]$,*

$$p(H_1)\, e^{\epsilon_-(H-H_1)} \le p(H) \le p(H_1)\, e^{\epsilon_+(H-H_1)}, \tag{G.6}$$

*so $R(H) = 1 - p(H)$ is trapped between two shifted exponentials. Similarly, if $\lambda_- \le \Lambda(H) \le \lambda_+$ on $[H_1, H_2]$, then $\mathrm{logit}\,p(H)$ is trapped between two affine functions of $H$ and $p(H)$ is trapped between two logistic curves.*

*Proof.* By (G.4), $\mathrm{d}\log p/\,\mathrm{d}H = \mathcal{B}(H)$. Integrating the bounds $\epsilon_- \le \mathcal{B}(H) \le \epsilon_+$ from $H_1$ to $H$ yields $\epsilon_-(H - H_1) \le \log\frac{p(H)}{p(H_1)} \le \epsilon_+(H - H_1)$, which is equivalent to (G.6) after exponentiating. The logistic bounds follow by the same argument applied to $\mathrm{d}(\mathrm{logit}\,p)/\,\mathrm{d}H = \Lambda(H)$ from (G.3). □

### G.2 A SOLVABLE HEAD-TAIL MANIFOLD

To make (G.3)–(G.4) concrete beyond the terminal limit, it is useful to study a low-dimensional invariant family that matches our numerical surrogates. Fix $K \ge 2$, set $n = K - 1$, and consider the symmetric submanifold

$$y(\tau) = \Big(1 - m(\tau),\ \tfrac{m(\tau)}{n}, \ldots, \tfrac{m(\tau)}{n}\Big), \qquad m \in \big(0, \tfrac{n}{n+1}\big), \tag{G.7}$$

together with the uniform bad block

$$z(\tau) = \Big(\tfrac{1}{M}, \ldots, \tfrac{1}{M}\Big). \tag{G.8}$$

Both manifolds are invariant by symmetry. On them, the residual good-tail mass $m$ obeys the exact cubic ODE

$$\frac{\mathrm{d}m}{\mathrm{d}\tau} = -m(1-m)\Big(1 - m - \tfrac{m}{n}\Big). \tag{G.9}$$

The entropy is explicit:

$$H(m) = h_{\mathrm{bin}}(m) + m \log n, \qquad \frac{\mathrm{d}H}{\mathrm{d}m} = \log \frac{n(1-m)}{m}. \tag{G.10}$$

Combining (G.9) and (G.10) yields an explicit covariance term

$$\mathcal{C}(m) = m(1-m)\Big(1 - m - \tfrac{m}{n}\Big) \log \frac{n(1-m)}{m}. \tag{G.11}$$

Since (G.8) gives $t_2 \equiv 1/M$, the bad-mass logit also has a closed form along this manifold:

$$\begin{aligned} \operatorname{logit} p(m) = \operatorname{logit} p_0 &+ \frac{M+1}{M} \log \frac{m}{m_0} \\ &+ \frac{M+n}{M} \log \frac{1-m}{1-m_0} \\ &- \frac{M+n+1}{M} \log \frac{1-m-m/n}{1-m_0-m_0/n}. \end{aligned} \tag{G.12}$$

Equations (G.10) and (G.12) give a complete non-asymptotic parametric description of $p(H)$ beyond the tail.

## H Sensitivity to effective bad-block dimension

We isolate the effect of the effective bad-block dimension $M$ by using a uniform bad block. The simulation ensemble in this section reuses the good-block initialization families from Appendix B.2 and replaces only the bad block by a uniform distribution over $M$ coarse-grained bad modes. In this case the bad-block contribution has

$$A_{-}(U(I)) = \exp(-\tau(I)/M), \qquad \mathcal{T}(I) = 1/M,$$

so the exact ODE solution can be written without numerical integration as

$$\operatorname{logit} p(I) = \operatorname{logit} p_0 - \log A(I) - \frac{\tau(I)}{M}.$$

The two bridge hazards are therefore

$$\Lambda(H) = \frac{d}{dH} \operatorname{logit} p = \frac{\mathcal{S}(I) + 1/M}{\mathcal{C}(I)}, \qquad \mathcal{B}(H) = \frac{d}{dH} \log p = (1 - p(H))\Lambda(H).$$

We compare $M \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ by matching normalized median hazard curves to the empirical curves over the observed accuracy range $a \in [0.05, 0.45]$. Small uniform $M$ should not be read as a literal estimate of the number of incorrect traces. It represents the concentrated end of a spectrum of effective bad-block dimensions. As $M_{\mathrm{eff}}$ grows, the bad-block entropy term $pH_{\mathrm{bad}}$ becomes more important, and the full-entropy curve can deviate from a single global logistic graph. This sweep therefore delineates the regime in which the reduced hazard diagnostics give a good finite-window approximation, rather than fitting a physical value of $M$.

**Effective, not literal, bad support.** The parameter $M$ should not be interpreted as the raw number of possible incorrect token sequences. The raw bad space in mathematical reasoning or coding is enormous. In the reduced model, bad modes are coarse-grained failure families, and the dynamics depend on their effective collision dimension through

$$\mathcal{T} = \|z\|_2^2, \qquad M_{\mathrm{eff}}^{(2)} := \frac{1}{\|z\|_2^2}.$$

Thus a large raw support can have small collision dimension when bad mass concentrates on a few families. In a uniform block, $M_{\mathrm{eff}}^{(2)} = M$; otherwise $M_{\mathrm{eff}}^{(2)}$ changes along the trajectory and is not interchangeable with the active-support count $M$ in the terminal exponent. Collision dimension alone also does not determine the bad-block Shannon entropy. In rollout logs, a practical estimate would first group incorrect completions into coarse failure families, for example by exact normalized answer, verifier failure signature, or an embedding/trace cluster, then set $\widehat{z}_c$ to the conditional frequency of bad rollouts in family $c$ and report $\widehat{M}_{\mathrm{eff}}^{(2)} = 1/\sum_c \widehat{z}_c^2$. For a finite sample, a pair-count estimator of collision probability can reduce the plug-in bias, with uncertainty reported separately. We therefore do not use the public aggregate curves to infer a literal bad-support size. The uniform-$M$ sweep is only a sensitivity analysis showing how the hazard diagnostics change as the bad block moves from concentrated to diffuse.

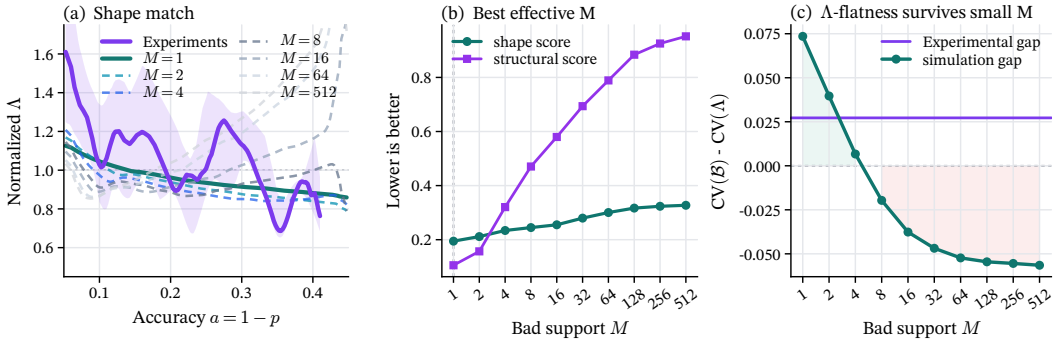


Figure 17: Effect of the effective bad-block dimension $M$ on the logistic-hazard flatness diagnostic. Panel (a) compares normalized logit-side hazard curves from exact uniform-$M$ simulations against the empirical median. Panel (b) shows two matching scores across $M$; the scores summarize how the reduced hazard diagnostics change with effective support. Panel (c) shows the logistic flatness advantage, measured by $\mathrm{CV}(\mathcal{B}) - \mathrm{CV}(\Lambda)$; positive values favor the sigmoid/logistic bridge.

**Matching scores.** For each candidate bad-support size $M$ we compare the exact uniform-$M$ simulation against the empirical ensemble on the observed accuracy window $a \in [0.05, 0.45]$ with two complementary scores. The first, the *shape score*, measures how well the simulation reproduces the *shape* of the empirical normalized hazard curves:

$$S_{\mathrm{shape}}(M) \;=\; \tfrac{1}{2}\big(\mathrm{RMSE}_a\big[\bar{\Lambda}_{\mathrm{real}}(a) - \bar{\Lambda}_M(a)\big] \;+\; \mathrm{RMSE}_a\big[\bar{\mathcal{B}}_{\mathrm{real}}(a) - \bar{\mathcal{B}}_M(a)\big]\big)\,, \tag{H.1}$$

where $\bar{\Lambda}_{\mathrm{real}}$, $\bar{\Lambda}_M$, $\bar{\mathcal{B}}_{\mathrm{real}}$, and $\bar{\mathcal{B}}_M$ denote the medians of the corresponding per-curve normalized hazards over the real ensemble and the simulated uniform-$M$ ensemble respectively (each curve being normalized by its mean hazard on the same window before aggregation). This score is scale-free and captures geometric agreement but is blind to whether the simulation preserves the qualitative $\mathrm{CV}(\Lambda) < \mathrm{CV}(\mathcal{B})$ inequality that motivates the sigmoid bridge. The *structural score* therefore combines the two shape errors with three curve-level statistical fingerprints, the error on the mean CV-gap $\overline{\mathrm{CV}(\mathcal{B}) - \mathrm{CV}(\Lambda)}$, the error on the fraction of curves with $\Lambda$ flatter than $\mathcal{B}$, and the error on the mean logit-linearity advantage $\overline{\Delta R^2} = \overline{R^2_{\mathrm{logit}} - R^2_{\mathrm{exp}}}$. Because these five errors have very different natural scales, each is min–max normalized to $[0, 1]$ across $M$, and then averaged:

$$S_{\mathrm{struct}}(M) = \tfrac{1}{5} \sum_{c \in \mathcal{C}} \frac{e_c(M) - \min_{M'} e_c(M')}{\max_{M'} e_c(M') - \min_{M'} e_c(M')},$$
$$\mathcal{C} = \{\mathrm{RMSE}_\Lambda,\, \mathrm{RMSE}_{\mathcal{B}},\, e_{\mathrm{gap}},\, e_{\mathrm{frac}},\, e_{\Delta R^2}\}. \tag{H.2}$$

A component with zero range across $M$ contributes zero to the normalized score. Lower is better for both scores. They are complementary because the shape score rewards pointwise agreement of the median curves while the structural score additionally rewards preserving the flatness and linearity signatures that are the stated comparison between the logistic and shifted-exponential approximations. We rank the uniform-$M$ simulations by $S_{\mathrm{struct}}$ and report $S_{\mathrm{shape}}$ alongside it; this ranking does not identify a physical support size.

Figure 17 summarizes the main sweep over effective bad-block dimensions, while Figure 18 provides the supplemental diagnostics used to check robustness.

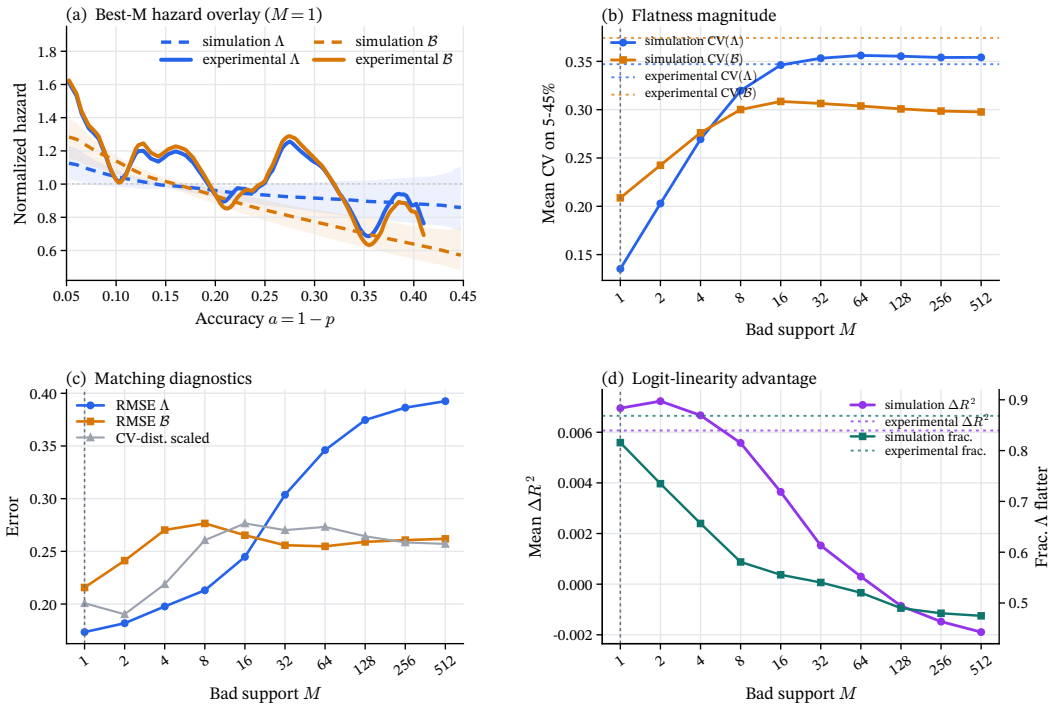


Figure 18: Additional $M$-sweep diagnostics. The uniform-$M$ simulations illustrate how normalized hazard shapes vary with effective bad-block dimension. Larger $M$ values can match some noisy CV magnitudes, but they progressively reduce or reverse the $\Lambda$-flatness advantage.